1-1-2014

Peto's Paradox and the Evolution of Cancer Suppression

Aleah Fox Caulin
University of Pennsylvania, aleah.caulin@gmail.com

Peto's Paradox and the Evolution of Cancer Suppression

Abstract
In order to successfully build and maintain a multicellular body, somatic cells must be constrained from proliferating uncontrollably and destroying the organism. If all mammalian cells were equally susceptible to oncogenic mutations and had identical tumor suppressor mechanisms, one would expect that the risk of cancer would be proportional to the body size and lifespan of a species. This is because a greater number of cells and cell divisions over a lifetime would increase the chance of accumulating mutations that result in malignant transformation. Peto's paradox is the clash between the theory that cancer incidence should increase with body size and lifespan, and the observation that it does not. In this thesis, I present the first comprehensive survey of empirical evidence across mammals in support of Peto's paradox in addition to computational models that explore the numerous hypotheses that may help resolve the paradox. I provide a detailed examination of tumor suppression in African elephants (*Loxodonta africana*) and show that the genome contains redundant copies of the tumor suppressor gene *TP53*. I give evidence that these redundant copies are actively transcribed and also observe an increased apoptotic response after exposure to ionizing radiation, which may be linked to the expression of these genes. Few genomes of large, long-lived organisms are currently available, which motivated my work to provide the sequence and de novo assembly of the humpback whale (*Megaptera novaeangliae*) genome. In this genome, I discovered a set of tumor suppressor genes that have evolved at an accelerated rate along the whale lineage, which is suggestive of adaptation. Additionally, I find one gene that has undergone convergent evolution between the African elephant and the humpback whale. The overarching goal of my research is to gain a better understanding of how evolution has suppressed cancer in large, long-lived organisms in the hopes of ultimately developing improved cancer prevention in humans.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Shane T. Jensen

Second Advisor
Carlo C. Maley

Keywords
African elephant, cancer prevention, cancer suppression, evolution, humpback whale, peto's paradox

PETO'S PARADOX AND THE EVOLUTION OF CANCER SUPPRESSION

Aleah Fox Caulin

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

_____
Shane T. Jensen, Ph.D.
Associate Professor of Statistics, UPenn

Co-Supervisor of Dissertation

_____
Carlo C. Maley, Ph.D.
Associate Professor of Surgery, UCSF

Graduate Group Chairperson

_____
Maja Bucan, Ph.D.
Professor of Genetics, UPenn

Dissertation Committee:

Junhyong Kim, Ph.D., Professor of Biology, UPenn
Aurora Nedelcu, Ph.D., Professor of Biology, University of New Brunswick
Harold Riethman, Ph.D., Associate Professor of Mol. & Cellular Oncogenesis, Wistar Inst.
Li-San Wang, Ph.D., Associate Professor of Pathology & Laboratory Medicine, UPenn

PETO'S PARADOX AND THE EVOLUTION OF CANCER SUPPRESSION

ACKNOWLEDGMENTS

First and foremost I would like to sincerely thank my advisors, Dr. Shane Jensen and Dr. Carlo Maley. Their enthusiasm for science and creativity has been a constant source of motivation. I appreciate the close guidance I was given at that times I needed it most, as well as the freedom I was given to develop as an independent researcher. They have taught more than they realize about both science and life and I hope to maintain close relationships with them throughout my career.

My committee has also been a great resource and I'm lucky to have been given the time to discuss my research with a group of scientists who always had useful advice and interesting ideas to contribute. I am truly grateful for the time that they have invested in me and for all of their suggestions, which have greatly contributed to my research. The GCB faculty and staff have also supported me in so many ways and I want to thank them for being so accommodating with anything I needed. Hannah Chervitz deserves a very special thank you for keeping us all in line and being a constant source of help.

Over the years I have had the privilege of working with many amazing people. Kristin, Lauren, Kathleen and Rumen welcomed me into the lab and quickly became more than lab mates. I am so thankful we have remained in touch, and though we've all gone our separate ways, they have continued to answer my many questions. The members of "Maley Lab 2.0" (Trevor, Ruchira, Amy and Viola) provided a great support system to discuss both science and life. They always look out for me and whether I need help with lab work, someone to edit my paper, or just a friend at happy hour they are more than willing to fill the role. I am also very lucky to have pseudo-lab mates (Helen and Greg) who have kept me sane through the transitional times of the lab. Having bi-coastal labs can be a challenge, but thankfully Nick and Sameer have managed to keep in touch no matter where I was.

I have been introduced to some impressive researchers over the years, a few of which I have had the pleasure of collaborating with. Many thanks to Dr. Joshua Schiffman who managed to obtain fresh elephant blood for my project and has shown me that with enough enthusiasm and drive, anything is possible. In his lab, Ashley Chan has

spent hours running irradiation experiments for me and I cannot possibly thank her enough. Also, Dr. Nader Pourmand and his at UCSC who performed all of our sequencing were such a pleasure to work with, as was Dr. Ed Green who provided great advice on genome assembly.

I feel so privileged to have such a supportive and loving family. I owe my parents for fostering my love of science from a very young age and teaching me to embrace my inner geek. As my mom would say, they're my 'biggest cheerleaders', and always have been. I am also lucky enough to have a sister who is always willing to discuss science and help me with anything she can. I am forever grateful for my husband, Brian, who I truly cannot thank enough. He has not only put up with all of the little things that go along with being married to a graduate student, but he has been my biggest supporter throughout the years. He has looked at more gel images and sequence alignments without complaint than any non-scientist I know, and always was excited about results he only partially understood just because he knew it was important to me.

Finally, I want to say thank you to those who have funded my work. I was lucky enough to be awarded the Department of Energy Computational Science Graduate Fellowship (DE-FG02-97ER25308), for which I am incredibly grateful. It has provided me opportunities to work in the national labs on exciting projects and the computational resources that I was given access to were essential for the work presented in this thesis. I did not get through the past 5+ years on my own, and I am thankful for each person, many of whom I do not have room to list specifically here, that played a role in helping me get to where I am today.

ABSTRACT

PETO'S PARADOX AND THE EVOLUTION OF CANCER SUPPRESSION

Aleah F. Caulin

Shane T. Jensen

Carlo C. Maley

In order to successfully build and maintain a multicellular body, somatic cells must be constrained from proliferating uncontrollably and destroying the organism. If all mammalian cells were equally susceptible to oncogenic mutations and had identical tumor suppressor mechanisms, one would expect that the risk of cancer would be proportional to the body size and lifespan of a species. This is because a greater number of cells and cell divisions over a lifetime would increase the chance of accumulating mutations that result in malignant transformation. Peto's paradox is the clash between the theory that cancer incidence should increase with body size and lifespan, and the observation that it does not. In this thesis, I present the first comprehensive survey of empirical evidence across mammals in support of Peto's paradox in addition to computational models that explore the numerous hypotheses that may help resolve the paradox. I provide a detailed examination of tumor suppression in African elephants (*Loxodonta africana*) and show that the genome contains redundant copies of the tumor suppressor gene *TP53.* I give evidence that these redundant copies are actively transcribed and also observe an increased apoptotic response after exposure to ionizing radiation, which may be linked to the expression of these genes. Few genomes of large,

long-lived organisms are currently available, which motivated my work to provide the sequence and *de novo* assembly of the humpback whale (*Megaptera novaeangliae*) genome. In this genome, I discovered a set of tumor suppressor genes that have evolved at an accelerated rate along the whale lineage, which is suggestive of adaptation. Additionally, I find one gene that has undergone convergent evolution between the African elephant and the humpback whale. The overarching goal of my research is to gain a better understanding of how evolution has suppressed cancer in large, long-lived organisms in the hopes of ultimately developing improved cancer prevention in humans.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

A portion of the text and figures found in this thesis has been previously published. Large portions of Chapter 1, parts of the introduction in Chapters 2 and the general discussion about the algebraic model (Model 1) in Chapter 3 are reprinted from the publication Trends in Ecology and Evolution, Vol. 26(4), Aleah F. Caulin and Carlo C. Maley. *Peto's Paradox: evolution's prescription for cancer prevention*. pages 175-182, Copyright (2011), with permission from Elsevier. Additionally some of the suggestions for future work were previously proposed in this publication. Figures that have been reprinted are noted in the captions. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

CHAPTER 1: INTRODUCTION

"Evolution is a light which illuminates all facts,
a curve that all lines must follow."
(Teilhard de Chardin 1959)

Cancer is the second leading cause of death in the United States (Hoyert and Xu 2012). Billions of dollars have been invested in cancer research; however, despite these efforts, 33% of Americans are diagnosed with cancer in their lifetime and 25% of our population still dies from this disease (ACS 2013). In this thesis, I approach the problem of cancer from a non-traditional angle, using evolutionary theory as the foundation for my research. I focus heavily on comparative genomics to gain insight into how evolution has shaped cancer suppression mechanisms across species. The long-term goal of this study is to translate the knowledge we gain from non-model organisms to the clinic, in hopes of decreasing the lifetime cancer risk of humans and stopping cancer in individuals before it begins.

THE EVOLUTIONARY THEORY OF CANCER

Cancer is a consequence of multicellularity and a striking example of multi-level selection. The theory of cancer initiation and progression is deeply rooted in evolutionary and ecological concepts (Merlo et al. 2006). Cancer progresses through somatic evolution, whereby genetic and epigenetic instability generates fitness variation among

cells. Throughout an organism's lifetime, cells accumulate mutations, which can eventually lead to the initiation of a malignancy. Mutations arise from both endogenous and exogenous damage, in addition to errors in DNA synthesis that are not properly repaired. The population of somatic cells within a tumor satisfy the three necessary and sufficient conditions for natural selection (Nowell 1976):

1. Th*ere must be variation within the population.* A tumor is a heterogeneous population of cells with somatic genetic and epigenetic alterations.

2. *The variation must be heritable.* Genetic and epigenetic mutations are inherited by both daughter cells when a cell divides.

3. *There must be differential survival and reproduction (i.e. fitness).* In some cases, the genetic and epigenetic mutations provide cells with survival and/or reproductive advantages over other cells.

Genetic and epigenetic changes can result in the eight 'hallmarks of cancer', all of which provide a fitness advantage to cancer cells relative to the healthy somatic cells: (i) self sufficiency of growth-signals, (ii) insensitivity to anti-growth signals, (iii) evasion of apoptosis, (iv) sustained angiogenesis, (v) limitless replicative potential (i.e. stabilization of telomeres), (vi) immune system avoidance, (vii) modification of cell metabolism, and (viii) the ability to invade new tissue and metastasize (Hanahan 2000, Hanahan and Weinberg 2011). The somatic evolution that takes place within mutant cell populations can result in cancer (Nowell 1976, Merlo, Pepper et al. 2006). Understanding this process through an evolutionary perspective is essential not only to guide treatment, but

additionally, and perhaps more importantly, for discovering ways to effectively intervene in order to prevent the development of cancer all together.

Just as selection acts at the cellular level, selection also acts at the level of the organism. This has led to the evolution of tumor suppressor mechanisms, such as cell cycle checkpoints and apoptosis, which act as safeguards to prevent somatic mutations from propagating in the cell population within a multicellular organism (Bernstein et al. 2002). DNA damage sensing and repair are crucial for resolving mutations as they arise, while premature senescence and apoptosis act as the second line of defense when mutations cannot be sufficiently repaired (Kinzler and Vogelstein 1997, Campisi 2003). Maintaining the integrity of DNA is essential for all forms of life, from unicellular to multicellular organisms; however, additional mechanisms have evolved in multicellular organisms to enforce cooperation and eliminate selfish cells in order to prevent cancer (Domazet-Loso and Tautz 2010).

Limiting the replicative potential of cells is thought to be an important mechanism involved in suppressing tumorigenesis (Campisi 1997, Campisi 2001). When a cell undergoes senescence, it is removed from the pool of dividing cells and can no longer pass on mutations it may harbor. Cells have been found to senesce in response to a variety of stresses such as mutation, over-expression of an oncogene and changes in chromatin organization (Campisi 2005). Additionally, the complex dynamics of stem cells and their lineages (e.g. asymmetric divisions, transient amplifying cells leading to terminally differentiated cells, and having only a small number of stem cells) helps to maintain the integrity of each tissue while most mutations occur in evolutionary dead

ends (i.e. cells that will become terminally differentiated) (Cairns 1975, Clevers 2005, Greaves 2007).

Specific tissue architectures (e.g. intestinal crypts) can also aid in tumor suppression by providing a physical barrier that separates small groups of stem cells, protects them in a niche and requires progeny to differentiate and slough off (Gatenby et al. 2010). When mutant cells arise and are not eliminated by any of these mechanisms, the immune system may be able to detect them and remove them before a malignant tumor is formed (Shankaran et al. 2001). Though these mechanisms are not necessarily solely in place to suppress tumorigenesis, the need to repress cancer has provided a strong selective pressure to fine-tune these systems throughout approximately one billion years of evolution in multicellular eukaryotes (Graham 1992, Knoll et al. 2006).

PETO'S PARADOX

The challenge of suppressing somatic evolution (i.e. cancer) dramatically increases with larger bodies and longer lifespans. If all mammalian cells were equally susceptible to oncogenic mutations and had identical tumor suppressor mechanisms, one would expect that the risk of cancer would be proportional to the body size and lifespan of a species. A greater number of cells in larger animals, and a greater number of lifetime cell divisions in long-lived animals, should increase the chance of accumulating oncogenic mutations. Some evidence exists that this is true within species (Altman and Schwartz 1978, Albanes 1998, Nunney 2013); however, there is no indication that this relationship holds across species. It is well documented that carcinogenesis is an increasing function of age (Frank

2007), and larger organisms generally have longer lifespans (Speakman 2005), which further suggests that we should see increased cancer incidence in large, long-lived animals. Peto's paradox is the clash between the theory that cancer incidence should increase with body size and lifespan, and the observation that it does not (Figure 1) (Peto et al. 1975, Peto 1977, Caulin and Maley 2011, Roche et al. 2012).

Cancer rates across multicellular animals only vary by approximately two-fold even though the difference of size among mammals alone can be on the order of one million-fold (Leroi et al. 2003, de Magalhães and Costa 2009). The exact functional relationship between body size and expected cancer risk is unclear; however it is assumed to be an increasing function (Figure 1). In comparing laboratory rodents and humans, which differ in maximum lifespan by a factor of 40 and size by three orders of magnitude, about 30% of both rodents and humans develop cancer within their lifetime (Rangarajan and Weinberg 2003). The general explanation for this is that large, long-lived animals are more resistant to carcinogenesis than small, short-lived animals (Dawe et al. 1969, Cairns 1975, Peto, Roe et al. 1975, Peto 1977, Graham 1992, Roche, Hochberg et al. 2012). However, how they accomplish this resistance has yet to be established.

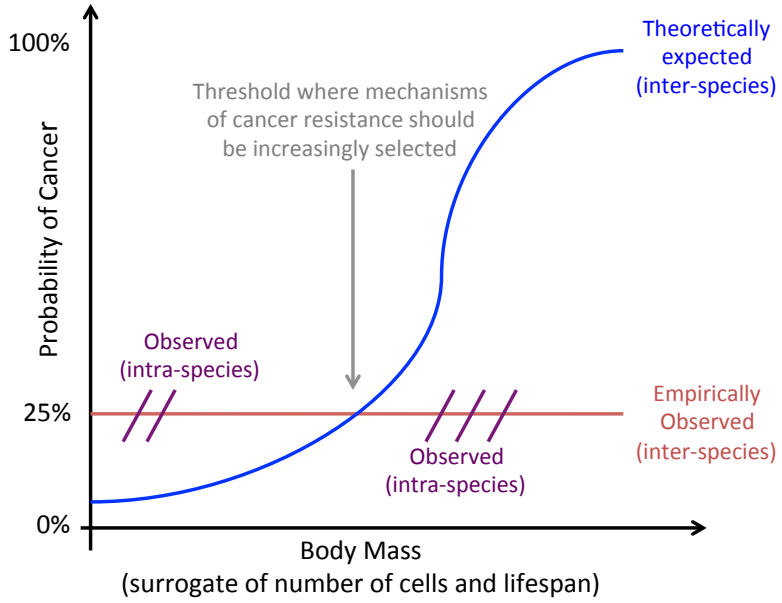


Figure 1. Summary of the current knowledge of Peto's paradox. Theory would predict that cancer incidence would increase with body size (blue curve); however across species the average probability of cancer is roughly constant (red line). Within a species we observe that body mass is associated with cancer risk (purple). This figure is adapted from Roche et al. 2012.

In the seminal paper bringing attention to this paradox, Sir Richard Peto noted: "in the evolutionary diversification of mammals, enormous changes in cellular susceptibility to oncogenesis have developed. These may be nicely illustrated by comparing mice and men: a man has 1,000 times as many cells as a mouse…and we usually live at least 30 times as long as mice do. Exposure of two *similar* organisms to risk of carcinoma, one for 30 times as long as the other would give perhaps 30^4 or 30^6 (i.e. a million or a billion) times the risk of carcinoma induction per epithelial cell. Are our stem cells really, then, a billion or a trillion times more 'cancer proof' than murine stem cells…Why don't we all die of multiple carcinomas at an early age? Presumably

some concomitant of our evolved ability to grow big and to live for three score years and ten is involved" (Peto 1977). Selection pressure for cancer suppression is stronger in species that are larger and live longer, thus humans have evolved more effective cancer suppression than mice, and larger animals, such as elephants and whales, have evolved more effective cancer suppression than humans. Understanding the molecular mechanisms involved in the evolution of tumor suppression could suggest new methods of cancer prevention in humans.

THE NEED AND POTENTIAL FOR CANCER PREVENTION

Cancer has proven difficult to cure. Since former U.S. president Richard Nixon declared the "War on Cancer" over 40 years ago, little progress has been made on reducing lifetime risk of cancer and increasing survival rates for patients with late stage diagnoses (Etzioni et al. 2003, ACS 2013). The majority of cancer research focuses on treatment rather than prevention and this often leads to the recurrence of tumors that are resistant to therapy. With 10^9–10^{12} cells in a tumor and perhaps 10^5 mutations (Bielas et al. 2006, Sjoblom et al. 2006, Greenman et al. 2007, Mardis et al. 2009), it appears that in many cases therapy selects for a resistant clone (Merlo, Pepper et al. 2006). Increasingly, attention is turning to cancer prevention so as to avoid this scenario entirely.

A proven strategy in drug development has been to seek natural products that have been honed by millions of years of evolution to generate the desired effect (Newman and Cragg 2007). Peto's paradox suggests that the same approach can be used in cancer research. If large, long-lived animals, such as the whales, have evolved

mechanisms capable of suppressing cancer 1,000 times better than humans, they may hold the key for cancer prevention in humans.

HYPOTHESES TO RESOLVE PETO'S PARADOX

Limited research efforts have been focused on resolving Peto's paradox. However, there are many hypotheses that might explain how organisms could overcome the burden of cancer despite an increased number of cells and extended lifespan. Some have been previously proposed (Totter 1980, Nunney 1999, Hahn and Weinberg 2002, Leroi, Koufopanou et al. 2003, Nagy et al. 2007, Klein 2009, Roche, Hochberg et al. 2012) and other mechanisms that we outline below are novel, to the best of our knowledge. Large bodies and long lifespans have evolved independently along multiple lineages; therefore, we would not expect that all large, long-lived animals have evolved the same mechanism(s) to suppress cancer, unless the suppression stems from an innate characteristic which they all share. Differences in diet and carcinogenic exposures (including pathogens, which are currently known to be associated with 15% of human cancers (zur Hausen 1999)) are unlikely explanations because there are many-fold differences in size between organisms with similar environments (e.g., dolphins and whales) and similar diets (e.g., elephants and mice are both herbivores). Here we present some possible mechanisms that might have evolved to reduce the expected correlation between body size, lifespan and cancer risk.

Lower Somatic Mutation Rates

If large animals have lower somatic mutation rates per cell generation, then more cell divisions would need to occur, compared to smaller animals, in order for a cell to acquire the necessary mutations to become malignant. Mutation rate is a function of the error rate and the rate at which these errors are repaired. A lower somatic mutation rate could be achieved through a number of mechanisms including better DNA damage detection and repair mechanisms, or better elimination of mutated cells. It appears that at least for the liver, the mutation rates per cell division are comparable between mice and humans (Leroi, Koufopanou et al. 2003), though more advanced methods to measure somatic mutation rates *in vivo* are needed to fully explore this hypothesis.

Redundancy of Tumor Suppressor Genes

Tumor suppressor genes (TSGs) are genes that increase the chance of progression to cancer when they are inactivated or deleted. Computational models have shown that cells of larger animals should have less tolerance for TSG inactivation, given that a phenotype occurs when both alleles are mutated (Roche et al. 2013). Added redundancy of tumor suppressor genes could suppress cancer in large animals by requiring more mutations occur to induce a malignant phenotype (Nunney 1999, Leroi, Koufopanou et al. 2003), and therefore making cells more tolerant to some TSG mutation (Figure 2).

The number of pathways that need to be inactivated to induce malignancy, typically through the mutation of TSGSs, differs between species with transformational

resistance thought to be highest in whales and lowest in mice (Lichtenstein 2005). For instance, the transformation of fibroblasts requires that 6 signal pathways be affected in humans, compared to only 2 in mice (Rangarajan et al. 2004). In support of the possibility of redundant TSGs, transgenic mice that contain an extra copy of *p53* (including its regulatory elements) gain an increased resistance to cancer and show no signs of premature aging (García-Cao et al. 2002). Similar results have been reported for p16 and ARF (Matheu et al. 2004). If large, long-lived animals evolved redundant copies of tumor suppressors, it might explain how they are not more prone to cancer any more so than smaller animals.

Redundancy could also be manifested in terms of TSG expression. Many tumor suppressor genes are tissue specific (Payne and Kemp 2005). Cells of larger species could have evolved expression patterns such that in any given cell more TSGs are expressed compared to smaller, shorter-lived animals, even though there might be the same number of TSGs in the genome. Cells could alter their expression of TSGs via epigenetic changes, introduction of transcription factor binding sites or non-coding RNA. This hypothesis would predict that large animals would have more ubiquitously expressed TSGs than smaller species.

Elimination of Proto-oncogenes

A complementary solution would be to eliminate a set of proto-oncogenes from the genomes of large, long-lived organisms. If there were fewer proto-oncogenes or proto-

oncogenic pathways that could generate the phenotypes necessary for cancer, there would be a reduced likelihood of cancer (Figure 2). This is supported by an experiment demonstrating that *Hras1* null mutant mice develop significantly fewer papillomas than wild type mice (Ise et al. 2000). This option might be constrained by selective pressures on the remaining pathways to produce the adaptive phenotypes that had been encoded in the deleted pathway. Proto-oncogenes serve specific cellular functions; so eliminating them could be deleterious for other reasons.

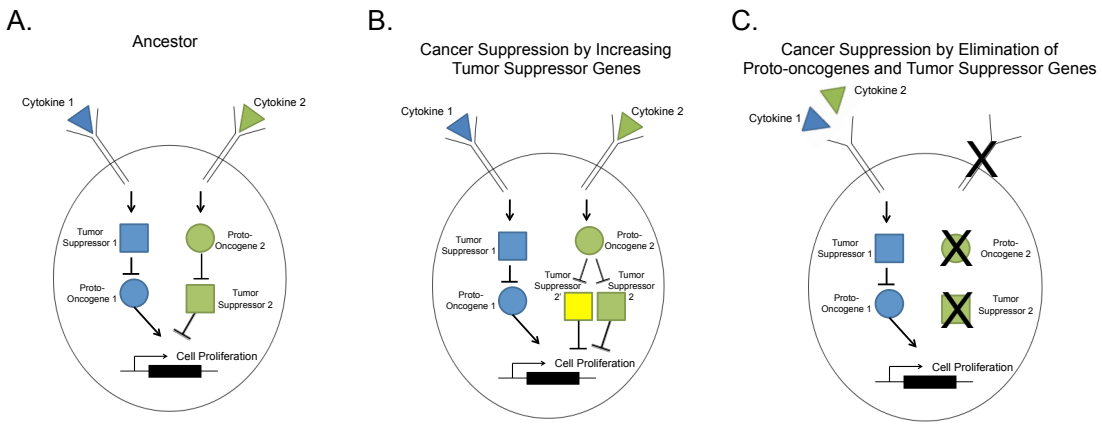


Figure 2. Mechanisms of cancer suppression. Assume that the ancestor of a large, long-lived organism has two pathways initiated by cytokines (triangles) such that if either one is disrupted the result is a hallmark of cancer. The example shown is for increased cell proliferation but could apply to any hallmark (A). A large organism could decrease its risk of cancer by evolving redundant copies of tumor suppressor genes (squares) (B) or by removing proto-oncogenes (circles) and tumor suppressor genes to eliminate an entire pathway (C) so that there are fewer carcinogenic loci in the genome that are vulnerable to mutation. This figure is reprinted from the publication Trends in Ecology and Evolution, Vol. 26(4), Aleah F. Caulin and Carlo C. Maley. *Peto's Paradox: evolution's prescription for cancer prevention*. pages 175-182, Copyright (2011), with permission from Elsevier.

Lower selective advantage of mutant cells

Another hypothesis to resolve Peto's paradox proposes that the fitness effect of specific mutations may differ in larger animals. For example, a haploinsufficient gene in mice could be completely recessive in a larger animal, requiring mutations to occur on both alleles in order to gain a selective advantage over neighboring cells during carcinogenesis in the larger species (Leroi, Koufopanou et al. 2003). This would decrease the possibility that mutations at this locus would contribute to progression towards cancer, which has been observed in a tissue-specific manner. The tumor suppressor *Trp53* (the mouse homolog to human *TP53)* usually requires both alleles of the gene to be null in order to see a mutant phenotype; however, in some tissues, *Trp53* is haploinsufficient and losing only one allele produces a phenotype in mice (Payne and Kemp 2005). If all cancer-associated genes required both alleles to be mutated in order to produce a phenotype, it would double the number of mutations required for a malignancy.

More sensitive or efficient apoptotic processes

The apoptotic propensity of cells might differ between large and small organisms. Cells from large bodies could be more sensitive to DNA damage or the activation of an oncogene and thus would be more apt to apoptose (Klein 2009). Support for this hypothesis comes from observations of human and mouse cell culture experiments. When human cells are exposed to methylating agents, many die via apoptosis triggered by the methylation damage. A much higher percentage of mouse cells survive and continue dividing regardless of the damage inflicted by the treatment (Humbert et al. 1999).

Apoptosis due to DNA damage eliminates the damaged cell from the population instead of repairing the DNA and possibly propagating remaining mutations in the tissue. However, there is likely a trade-off between apoptosis preventing cancer and accelerating aging due to depletion of the stem cell pool (Tyner et al. 2002).

Increased sensitivity to contact inhibition

Additionally, selfish cellular proliferation can also be suppressed by signals from the microenvironment (Klein 2009). For example, cell contact inhibition has been noted to differ between human, mouse and naked mole-rat (*Heterocephalus glaber*) cells. In culture, naked mole-rat cells stop dividing at much lower densities than human and mouse cells due to the early activation of the p16 pathway which results in hypersensitivity to contact inhibition (Seluanov et al. 2009). Although naked mole-rats and mice are small animals, the former live significantly longer than the latter (28 years (Buffenstein and Jarvis 2002) versus 4 years (Turturro et al. 1999)). In 250 necropsies of naked mole-rats that died in captivity, zero had cancer (Buffenstein 2005). Hypersensitivity to contact inhibition might have evolved to suppress cancer allowing the naked mole-rat to live longer, though this cellular response has only been verified *in vitro* (Seluanov, Hine et al. 2009). Similar signals for early cell senescence may be triggered in large, long-lived organisms to inhibit uncontrolled proliferation, and decrease the risk of tumorigenesis.

Shorter Telomeres

Telomere length appears to be a fundamental check on the proliferative capacity of cells (Monaghan 2010). Telomeres shorten with every cell cycle and when they become too short to protect the chromosomes' ends, the cell senses those ends as DNA double strand breaks, usually leading to apoptosis (d'Adda di Fagagna et al. 2003, Shay and Wright 2010). Even though stem cells express telomerase, which helps to rebuild telomeres, they generally do not express enough to prevent telomere shortening due to proliferation (Shay and Wright 2010). Within rodents, repression of telomerase activity coevolved with increased body mass, resulting in decreased expression in larger species (Seluanov et al. 2007). However, the long-lived, cancer-resistant naked mole-rat does express telomerase and has continuously proliferating cells, suggesting that larger bodies are a greater risk factor for cancer than longevity (Seluanov, Chen et al. 2007, Seluanov et al. 2008). We hypothesize that large, long-lived animals might have shorter telomeres (or erode them faster) than smaller animals, limiting the number of times their cells can divide and reducing opportunities to accumulate carcinogenic mutations, which could be assayed *in vitro.*

Different tissue architecture

Alternatively, tumor suppression mechanisms can evolve at the level of the tissue as opposed to within individual cells. Changes in tissue architecture could influence the frequency of cancers by altering the way cells are compartmentalized and/or the dynamics of the tissue (Leroi, Koufopanou et al. 2003). Most tissues are comprised of

small proliferative units, e.g. the crypts of the intestines. It has been proposed that this hierarchical structure is a crucial cancer prevention mechanism (Gatenby, Gillies et al. 2010). Since differentiating cells are evolutionary dead-ends because they will stop dividing and eventually slough off, the effective population size of a somatic tissue depends mainly on the number and dynamics of stem cells (though a mutation which disrupts differentiation in a non-stem cell might also generate a carcinogenic cell lineage) (Michor 2007). Under a model of "serial differentiation" it is possible to increase the number of cells and the amount of cell turnover without increasing the number or proliferative activity of somatic stem cells, simply by adding non-stem stages (Pepper et al. 2007). Altering the number of stem cells, the crypt density or the dynamics of differentiation and division could enhance the tissue's ability to prevent malignant transformation.

More effective immune system

The immune system has been found to play a role in preventing tumorigenesis and provides yet another possible resolution to Peto's paradox. Immune system efficiency against virus-associated cancers might also account for some differences observed in cancer rates within people (Klein 2009), but this could apply to non-viral cancers as well. Initially, tumors are likely to be immunogenic. When mice are treated with carcinogens, tumorigenesis is delayed by immune system surveillance (Koebel et al. 2007). However, as the tumor co-evolves with the immune system, tumor variants that go undetected are selected (termed "immunoediting") (Pawelec et al. 2010). "Chronic antigenic stress" can

result in exhaustion of the immune system leading to ineffective surveillance, similar to observations of chronic viral infections (Pawelec, Derhovanessian et al. 2010). Large, long-lived organisms might have improved immune surveillance to eliminate neoplastic cells before they become malignant.

Less Reactive Oxygen Species due to Lower Basal Metabolic Rate

A lower somatic mutation rate could also be a result of metabolism. Reactive oxygen species (ROS) are byproducts of metabolism and can cause DNA damage thought to contribute to aging and cancer (Wiseman and Halliwell 1996, Hoeijmakers 2009, Sedelnikova et al. 2010). The rate at which ROS are produced in a cell is a function of the basal metabolic rate (BMR) (Ku et al. 1993). BMR per unit mass (mass-specific BMR) is proportional to $M^{-1/4}$, where M is body mass (Savage et al. 2007) and has been shown to correlate with the amount of oxidative damage (Adelman et al. 1988). This leads to the prediction that cancer incidence, if largely caused by endogenous DNA damage, should also be proportional to $M^{-1/4}$ (Herman et al. 2011). Knocking out oxidative repair genes, and therefore allowing DNA damage from ROS to persist, results in increased tumor susceptibility in a variety of tissues, suggesting that DNA damage caused by ROS plays a causal role in tumor formation (Xie et al. 2004). Large animals should produce fewer ROS due to their lower mass-specific BMR and consequently have less endogenous DNA damage (Totter 1980).

There is some evidence linking cancer incidence to metabolism. The average BMR of women is 10% lower than that of men after adjusting for body mass,

composition, activity and age (Totter 1980) and women consistently have lower rates of cancer (ACS 2013). Naked mole-rats, for which spontaneous cancer has yet to be reported (Buffenstein 2005), have a mass-specific BMR that is much lower than expected given their size (de Magalhães and Costa 2009). Additionally, caloric restriction inhibits cancers in animal models and one explanation for this is that the decrease in caloric intake lowers the metabolic rate, therefore producing less ROS and subjecting the DNA to less endogenous damage (Longo and Fontana 2010). These observations could all be attributed to cells having less endogenous oxidative damage, which effectively results in a lower somatic mutation rate and a reduced cancer risk.

Formation of hypertumors

Nagy et al. have proposed an alternative hypothesis to resolve Peto's paradox: the formation of 'hypertumors' (Nagy, Victor et al. 2007). Natural selection within a tumor might favor cheater cells that take advantage of vasculature built by angiogenic cells. These cheaters could grow and parasitize the primary tumor by forming a 'hypertumor' that would reduce the overall fitness of the tumor and possibly cause the tumor to regress. Nagy et al. argue that lethal tumors must be drastically larger in larger animals, giving the hypertumor more time to evolve and force the parent tumor to become necrotic (Nagy, Victor et al. 2007). This model predicts that tumors in large organisms should be disproportionately more necrotic when compared to lethal tumors in smaller organisms, and that large animals should carry a burden of many non-lethal tumors (Nagy, Victor et al. 2007), though these predictions have yet to be tested experimentally.

Since large bodies have evolved multiple times, independently throughout evolution, each lineage may have evolved a different mechanism to overcome the problem of cancer. However, the last two hypotheses involving BMR and hypertumors would apply to all large, long-lived animals and perhaps provide some universal answer to Peto's paradox. For practical reasons, we have focused our research efforts on exploring the hypotheses of redundant tumor suppressor genes and the apoptotic response of cells to DNA damage.

OVERVIEW OF THESIS

In this thesis, I investigate the underlying evolutionary basis for Peto's paradox through the use of computational modeling and comparative genomics. In Chapter 2, I discuss currently published evidence for the existence of the paradox and present an empirical analysis of necropsy data to provide improved estimates of cancer incidence in non-human mammals and additional support for the paradox. In Chapter 3, I outline two models to explore the various hypotheses that have been proposed to explain Peto's paradox. I find that there are multiple biologically relevant solutions and choose to focus subsequent work on the copy number of cancer-associated genes which is discussed in detail in Chapter 4. In Chapter 5, I present my discovery of redundant copies of the tumor suppressor gene *TP53* in the African elephant genome and argue that these extra copies may be responsible for the hypersensitivity to ionizing radiation that I have analyzed. Chapter 6 explains the methods that I used to *de novo* assemble the genome of the humpback whale, which is one of the largest and most long-lived organisms that has been

sequenced to date. In Chapter 7 I provide an evolutionary analysis of the humpback whale genome (in addition to the African elephant) where I find evidence of convergent and accelerated evolution in specific tumor suppressor genes. In Chapter 8, I propose future experiments that will further our understanding of cancer suppression mechanisms. If we can understand how evolution has shaped these mechanisms to allow for large, long-lived organisms to suppress cancer, this knowledge could be applied towards improved methods of cancer prevention in humans.

CHAPTER 2: EMPIRICAL EVIDENCE IN SUPPORT OF PETO'S PARADOX

INTRODUCTION

Cancer incidence records for wild and captive animals are not well documented for most species as the majority of animals live and die unseen (McAloose and Newton 2009). This makes it difficult to directly compare records of humans and other animals, but it is still clear that cancer incidence does not scale with body size across species. For example, blue whales are thought to be the largest animals to have ever lived (Small 1971) and are three orders of magnitude larger than a human. However, if blue whales had one thousand times higher cancer incidence than humans, they would likely die before they were able to reproduce and the species would have quickly gone extinct (Lichtenstein 2005). The mere existence of whales suggests that is it possible to substantially reduce the rate of cancer incidence relative to humans.

Cancer death rates vary approximately two-fold across multicellular animals of drastically different size (Leroi, Koufopanou et al. 2003). When wild mice are raised in protected laboratory conditions 46% die of cancer (Andervont and Dunn 1962). Cancer is also responsible for about 20% of dog deaths (Morris and Dobson 2001), roughly 25% of human deaths in the United States (ACS 2013) and 18% of beluga whale deaths (Martineau et al. 2002), though the latter were living in a highly polluted estuary. Rare cases of cancer are discovered in blue whales, giving no evidence of elevated cancer risk in these species (Martineau, Lemberger et al. 2002, Newman and Smith 2006).

Additionally, a radio-epidemiologic study suggests that dinosaurs may have even had a decreased risk of cancer, relative to modern-day animals (Rothschild et al. 2003, McAloose and Newton 2009). Contrary to what we observe; if the probability of cancer scaled with size and longevity, mice should have the lowest incidence of these examples. However, cancer seems to account for approximately the same percentage of deaths, despite the size and lifespan of species (Figure 3).

Interestingly, within a species, size is associated with an increased cancer risk. In humans, having a leg length 3-4mm above average results in an 80% higher risk of non-smoking-related cancers (Albanes 1998). This trend is also seen in children and dogs. Children with bone cancers tend to be taller and osteosarcomas occur in large dogs 200 times more frequently than small and medium breeds (Altman and Schwartz 1978, Nunney 2013). There has likely not been enough time for larger dogs to evolve additional mechanisms to protect them from this increased risk and counteract the extreme artificial selection for size imposed by humans. This suggests that animals which evolved to be larger as a species developed mechanisms to offset the increased cancer risk associated with an increased number of cells. Conversely, above average individuals do not possess additional defenses compared to smaller organisms within their species, and therefore fall victim to cancer with greater probability.

CANCER INCIDENCE DATA OF ZOO MAMMALS

Necropsy data from animals in captivity confirms that cancer incidence does not increase with body size or lifespan across species varying orders of magnitude in both size and

lifespan. We compiled 14 years of necropsy data collected by the San Diego Zoo (Griner 1983) and counted the recorded instances of tumors for mammals. In over 830 necropsies across 36 mammals we found a total of 37 incidences of cancer, which is only an overall incidence of 4.5%. A previous study, which did not require a minimum of 10 necropsies per species, found that 2.75% of species had neoplasms at the time of necropsy (Effron et al. 1977). The highest rate of cancer in the data we analyzed was found in Tasmanian devils, though none of the cases were linked to the contagious facial cancer in that species. Tasmanian devils have very low genetic diversity, likely due to multiple population bottlenecks which are thought to be caused by disease (Guiler 1983). It is possible that a cancer-susceptibility allele increased in frequency due to these bottlenecks, leaving the remaining population highly vulnerable to malignant transformation.

A logistic regression model was fit to the data to test if body mass and maximum lifespan are good predictors of cancer incidence for each species. We found no evidence to support the predictive value of these characteristics (Figure 3). We also used a logistic regression to examine all combinations of mass, lifespan and mass-specific basal metabolic rate since these features are highly correlated; however, we still found no significant relationship with cancer incidence (Figure 4). If anything, the trend is toward a lower risk of cancer in animals with increased mass and lifespan. These analyses provide the first systematic evidence to support Peto's Paradox.

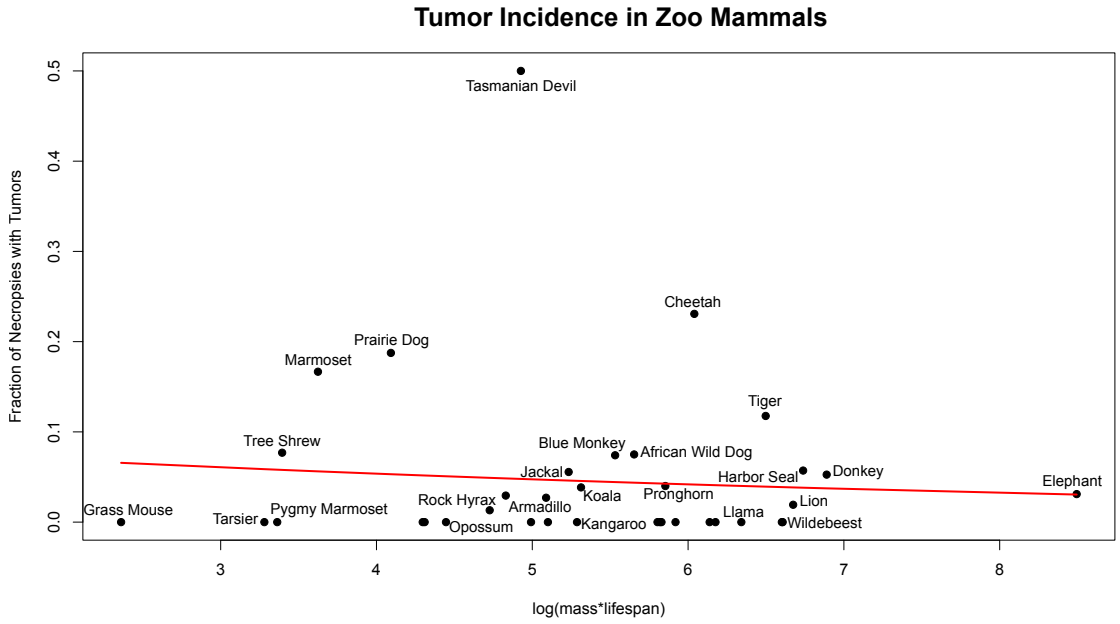


Figure 3. Tumor incidence in captive mammals. Cancer incidence does not increase with body size and lifespan. The product of mass and lifespan is not a good predictor of cancer incidence as shown by the logistic regression (model fit shown as red line). The elephant data point is from data collected from Elephant Encyclopedia database and is based off of 644 annotated deaths (Koehl 1995-2012). All other data are from necropsies at the San Diego Zoo (Griner 1983) and each point is supported by a minimum of 10 necropsies.

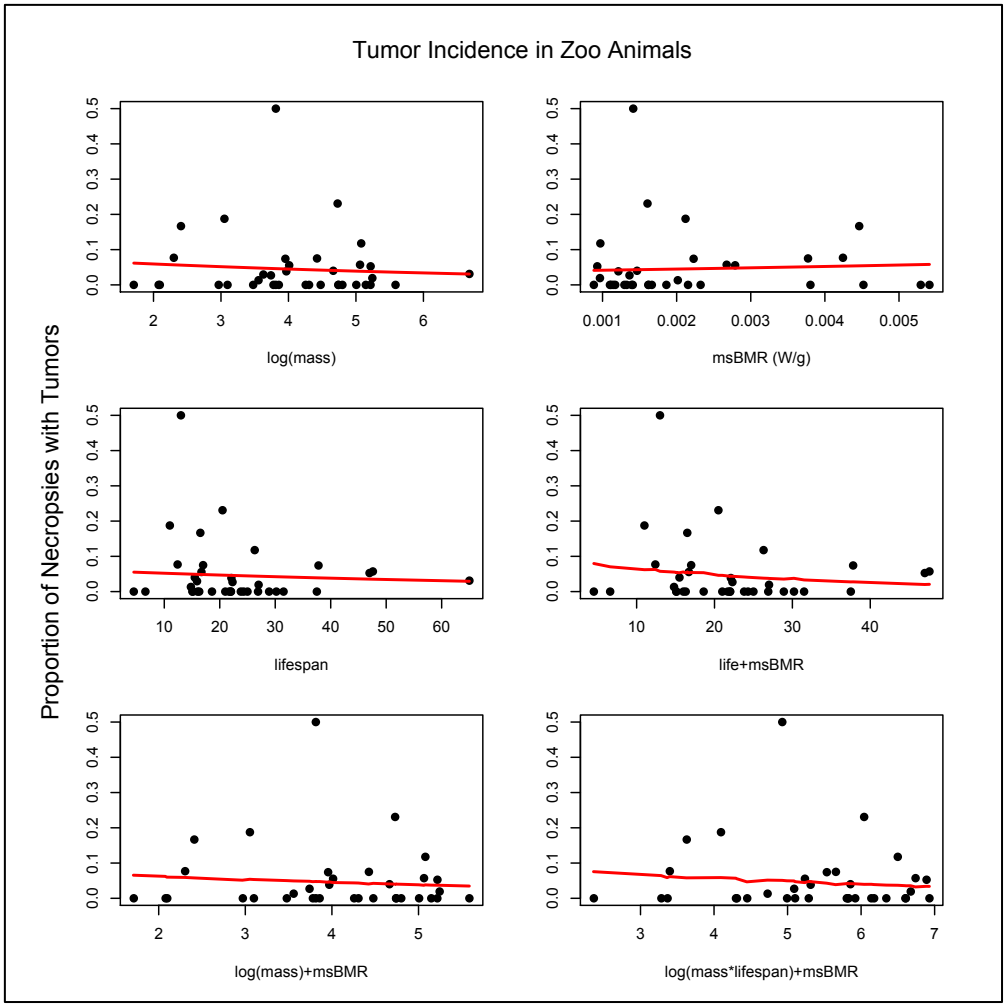


Figure 4. Logistic regression models for tumor incidence. Logistic regression models for tumor incidence in zoo animals show no significant correlation between tumor incidence and body size, lifespan, mass specific basal metabolic rate (msBMR) or any combination of those variables. The log(mass*lifespan) plot is shown in Figure 3.

AGE INCIDENCE AND LIFETIME RISK OF CANCER IN ELEPHANTS

Next we specifically investigated the cancer incidence in the largest extant terrestrial mammal, the elephant. If elephants had the same biology as humans, and cancer incidence scaled linearly with the number of cells and lifespan of a species, with 100-fold more cells and lifespans up to 65 years (de Magalhães and Costa 2009), which is more

than half the average human lifespan, elephants should get approximately 50-fold more cancers than humans. We analyzed data from the Elephant Encyclopedia (Koehl 1995-2012) on the cause of death for elephants in captivity in order to get an estimate of their age-incidence and overall lifetime risk of cancer (Figure 5). Out of 644 annotated deaths there were 20 cases of cancer/lethal tumors, resulting in a lifetime cancer incidence of 3.1%. The true cancer incidence is obscured by the fact that necropsies are not performed on all of the animals at time of death and elephants are frequently euthanized for reasons such as arthritis, aggression and injury. Many of the animals are euthanized because of "age related issues" which are unspecified and interfere with the cancer incidence data since this prevents many elephants in captivity from reaching the age at which they would naturally die. To get a more comprehensive estimate we calculated an inferred cancer incidence by assuming the same percentage of deaths with an unknown cause would be due to cancer as deaths with known causes (see Methods). Using this calculation, the lifetime cancer death rate in elephants in captivity only increased to 4.8%, compared to the 25% lifetime cancer mortality rates in humans in the United States (ACS 2013) and 13% worldwide (Ferlay et al. 2010). How are elephants suppressing cancer more effectively than humans?

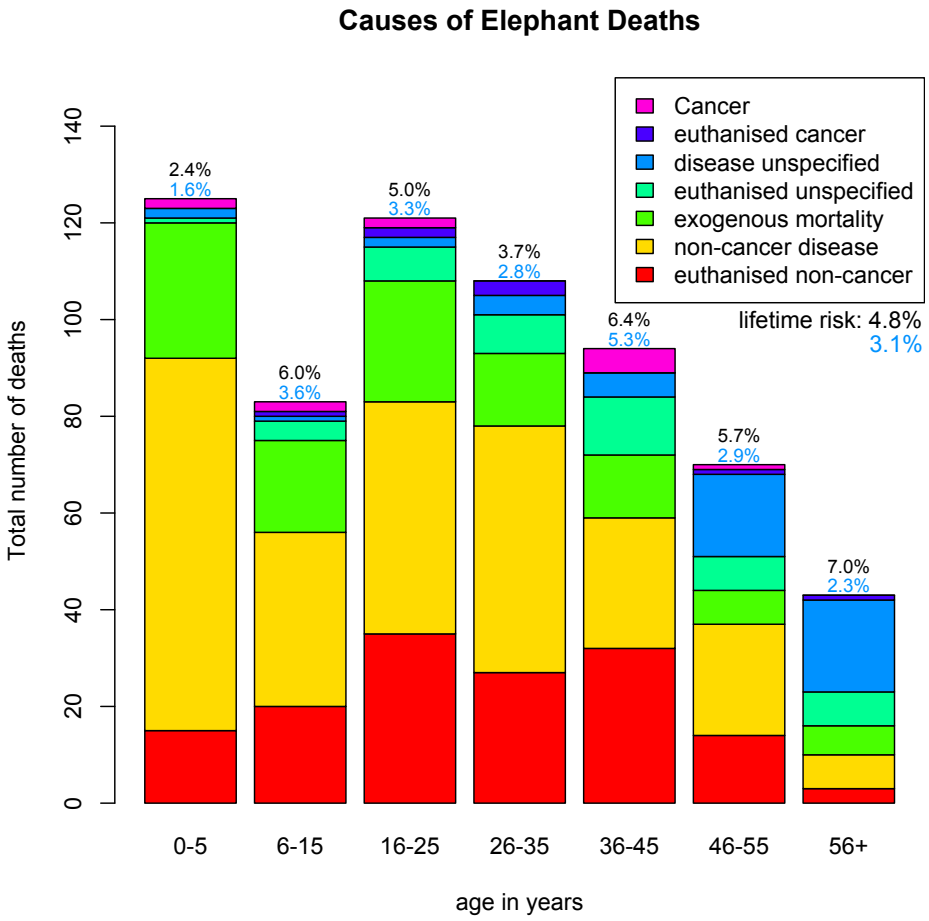


Figure 5. Cause of death in captive elephants. Approximately 3.1% of elephants die of cancer, which given their size and lifespan is much lower than we would expect. Because not all elephant deaths are well annotated, we calculated an estimated lifetime cancer risk, which increases the cancer incidence estimate to 4.8% (see Methods). The inferred cancer rates for each age group are shown in black and the observed rates are shown in blue.

METHODS

We compiled necropsy data collected by the San Diego Zoo (Griner 1983) to estimate cancer incidence in mammalian species. The analysis was limited to species with a minimum sample size of ten. This restriction gave a total of 832 necropsies across 36 species of mammals of which there were 37 reported cases of cancer and/or lethal tumors. Tumors that lacked full pathology reports, as well as cases noted as hyperplasias, were counted towards cancer incidence so as to not drastically underestimate the values. Adult body mass, maximum lifespan and mass specific basal metabolic rate (msBMR) data (Table 1) were collected from the AnAge database (de Magalhães and Costa 2009). We performed a logistic regression to determine if body mass, lifespan, msBMR or a combination of these variables was a good predictor of cancer incidence within a species. The logistic regression was done on both a log and linear scale, but plotted on a log scale to easily visualize the large range of masses.

Data on 644 elephants (both African (*Loxodonta africana)* and Asian (*Elephas maximus*)) were obtained from the online Elephant Encyclopedia database (Koehl 1995-2012). Elephants in the circus, at temples, and owned by private dealers were excluded from the analysis since treatment of these animals is often not held to the same standards as a zoo or sanctuary. Causes of death were divided into seven categories: cancer, euthanized because of cancer, non-cancer disease, euthanized for a reason other than cancer, unspecified disease, euthanized for an unspecified reason, and exogenous cause of mortality. Inferred cancer rates were calculated by assuming the same percentage of deaths with an unknown cause would be due to cancer as deaths with known causes. For

example, if cancer makes up *x* percent of deaths with a known cause, then we assume cancer is also responsible for *x* percent of the deaths with an unspecified cause (i.e. "disease unspecified" and "euthanized unspecified").

We used the following methodology to infer cancer deaths among the unspecified cases. The fraction of cancers reported in deaths with a specified disease is f_{dk} and the fraction of elephant euthanizations attributed to cancer is f_{ek}, where the subscript k represents 'known' and the *d* and *e* represent 'disease' and 'euthanized' respectively. The number of deaths from unspecified diseases that we infer to be cancer is equal to $f_{dk} \times N_{du}$, where N_{du} is the number of deaths caused by an unspecified disease. Similarly the number of unspecified euthanizations that we infer to be cancer is equal to $f_{ek} \times N_{eu}$, where N_{eu} is the number of euthanizations with no specified reason. We take the ceiling integer for each of these values as a conservative measure to not underestimate the cancer incidence. The inferred cancer rate is equal to $\frac{(f_{dk} \times N_{du}) + (f_{ek} \times N_{eu}) + C_{dk} + C_{ek}}{N}$, where C_{dk} and C_{ek} are the number of cancer cases in the known disease population and the known euthanized population respectively and N is the total number of elephant deaths.

Common Name	# necropsies	# tumors	Adult mass (Kg)	Maximum lifespan (yrs)	msBMR (W/g)
Striped Grass Mouse	13	0	0.05	4.5	0.00452
Philippine Tarsier	13	0	0.12	16.0	0.00381
Pygmy Marmoset	15	0	0.12	18.6	0.00541
Treeshrew	13	1	0.20	12.4	0.00424
Marmoset	18	3	0.26	16.5	0.00446
Squirrel Monkey	17	0	0.93	30.2	0.00529
Prairie Dog	16	3	1.13	11.0	0.00212
Fennec Fox	10	0	1.25	16.3	0.00232
Viriginia Opossum	11	0	3.00	6.6	0.00186
Rock Hyrax	76	1	3.60	14.8	0.00202
Parma Wallaby	34	1	4.25	15.9	unknown
Armadillo	74	2	5.50	22.3	0.00136
Racoon	15	0	6.00	21.0	0.00215
Tasmanian devil	18	9	6.50	13.0	0.00141
Darma Wallaby	14	0	6.50	15.1	0.00162
Tree Kangaroo	18	0	7.20	26.9	0.00114
Blue Monkey	27	2	9.00	37.8	0.00223
Koala	26	1	9.30	22.1	0.00121
Black-backed Jackal	18	1	10.25	16.7	0.00279
Hamadryas Baboon	18	0	18.00	37.5	0.00166
Collared peccary	16	0	20.20	31.5	0.00162
African Wild Dog	40	3	26.50	17.0	0.00377
Eastern Wallaroo	40	0	30.00	22.0	0.00111
Pronghorn	25	1	46.10	15.5	0.00147
Cheetah	13	3	53.50	20.5	0.00161
Red Kangaroo	15	0	55.00	25.0	0.00110
Capybara	13	0	55.00	15.1	0.00139
Cougar	11	0	63.00	23.8	0.00133
Reindeer	16	0	101.25	21.7	0.00141
Harbor Seal	35	2	115.00	47.6	0.00267
Tiger	17	2	119.70	26.3	0.00097
Llama	18	0	140.00	28.9	0.00130
Blue Wildebeest	25	0	164.50	24.3	0.00117
Donkey	19	1	165.00	47.0	0.00093
Lion	52	1	175.00	27.0	0.00097
Moose	13	0	386.00	22.0	0.00088

Table 1. Tumor incidence, mass, lifespan, and metabolic rate of zoo mammals. Necropsy data was collected by the San Diego Zoo over 14 years (Griner 1983). Body mass, maximum lifespan and mass specific basal metabolic rate (msBMR) were all obtained from the AnAge database (de Magalhães and Costa 2009).

DISCUSSION

Previous to this study, Peto's paradox lacked strong empirical evidence that cancer risk does not increase with body size or lifespan across species. Our analysis provides the first comprehensive survey across mammals in support of Peto's paradox. The mammalian species we investigated span five orders of magnitude in size and one order or magnitude in lifespan. They range from the striped grass mouse at 51g and living only 4.5 years, to the African elephant weighing 4,800Kg and living for approximately 65 years (de Magalhães and Costa 2009). We find no evidence of an increased cancer risk in larger, more long-lived animals. If anything, the regression lines show a trend of decreased incidence as the $log(mass \times lifespan)$ increases, though no regression trends were statistically significant in either direction. We observe this across all combinations of features that we tested. The model for the relationship between cancer incidence and msBMR looks like it goes in the opposite direction of all other models; however this is because the larger animals have a smaller msBMR.

Studies should continue to collect cause of death information for non-model organisms. Many of the species we analyzed had less than 20 specimens, but as these numbers increase we can gain a more comprehensive view of cancer prevalence across mammals and which species may harbor enhanced suppression compared to humans. These data suggest that there should no longer be a debate of whether or not Peto's paradox exists; however, the challenge remains to explain the driving forces of this paradox, which we investigate in later chapters.

CHAPTER 3: COMPUTATIONAL MODELS OF CANCER INCIDENCE

INTRODUCTION

Computational methods are frequently used to model cancer incidence (Beerenwinkel et al. 2007, Calabrese and Shibata 2010, Martens et al. 2011, Do et al. 2013). Typically these models address questions involving cancer initiation and progression in humans. In this work, we used previously developed models of colorectal cancer incidence to ask a new question: what values of the parameters governing cancer incidence allow for the age-incidence of cancer to be similar across body sizes that differ by orders of magnitude?

Our goal was to explain Peto's paradox using parsimonious models and gain insight into how evolution may have fine-tuned factors that greatly influence cancer risk, such as, mutation rate, cell division rate, and the number of hits required for carcinogenesis. We used two models, which are generally similar but differ in the dynamics of cell lineages. Both models maintain a constant population of size N and have non-overlapping generations, representing the cell dynamics of normal colon tissue. We first explored a simple algebraic model (Calabrese et al. 2004, Calabrese and Shibata 2010), which assumes that mutations accumulate over time at a fixed rate determined by the number of cell divisions. Given a population of N cells, the model allows one to calculate the probability of a cell having the required number of mutations (k) to initiate cancer after d divisions (Figure 6A).

The second model is based on a Wright-Fisher process and was previously used to examine progression from a benign polyp to a malignant colorectal tumor (Beerenwinkel, Antal et al. 2007). A Wright-Fisher process differs from the previous model because it allows for cell lineage death (Figure 6B). Given a population of N cells in generation t with identical fitness, each cell has equal probability ($1/N$) of being the parent of any single cell in generation $t+1$. With probability $\left(1 - \frac{1}{N}\right)^N$, a cell from generation t will not have any progeny in generation $t+1$ and thus that the lineage will be completely eliminated from future generations. This model also has a constant mutation rate per cell division and defines cancer to be the accumulation of k mutations in one cell.

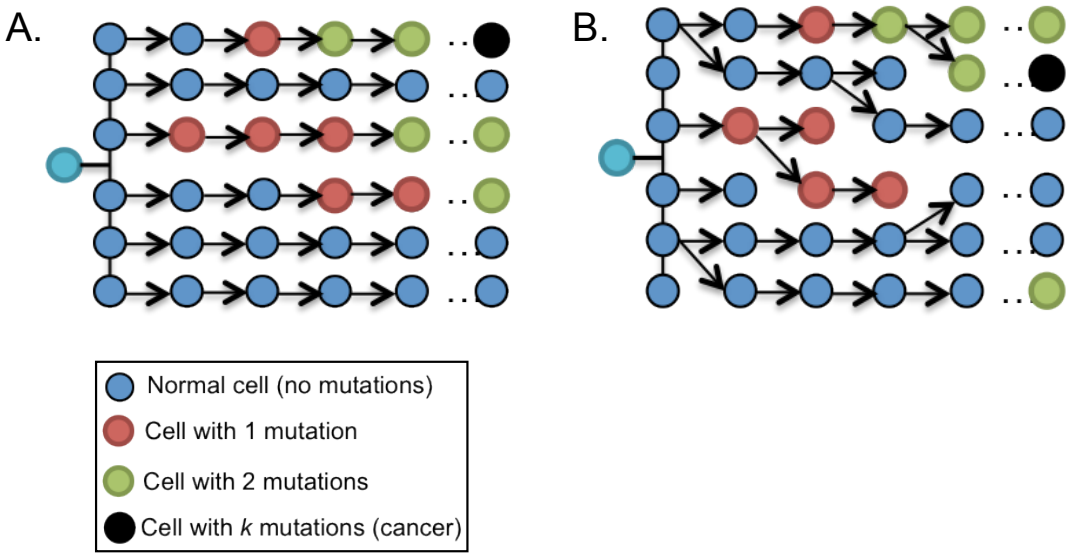


Figure 6. Model representation of cancer progression. In the algebraic model (A) (Calabrese and Shibata 2010), cell lineages accumulate mutations over time, which are passed on to their daughter cell in the next generation and there is no cell death. In the Wright-Fisher model (B) (Beerenwinkel, Antal et al. 2007), cells gain mutations over time, but each lineage has a chance of dying and being eliminated from the population. In both models, cancer occurs when a cell accumulates k mutations. The single light blue cell represents the zygote to show that all cells came from a single initial lineage.

MODEL 1: ALGEBRAIC MODEL OF CANCER INCIDENCE

Calabrese and Shibata devised a simple mathematical equation to express the probability of a human developing colorectal cancer given their age (Calabrese and Shibata 2010). Their equation produces results which closely match incidence data from the Surveillance, Epidemiology and End Results (SEER) database (SEER 2001). The probability of an individual developing colorectal cancer after a given number of stem cell divisions is

$$p = 1-(1-(1-(1-u)^d)^k)^{Nm}$$

where u is the mutation rate per gene per division, d is the number of stem cell divisions since birth, k is the number of rate limiting mutations required for cancer to occur, N is the number of effective stem cells per crypt and m is the number of crypts per colon (Calabrese and Shibata 2010).

The model also shows that the increased cancer risk observed in taller women in the SEER data set can be fit by simply increasing the parameter m to account for a larger colon (Calabrese and Shibata 2010). Using the same rationale, we varied the parameter m from 1.5×10^3 to 1.5×10^{10} to see how the total number of stem cells in the colon changes the lifetime (90 year) risk of developing colorectal cancer (Figure 7). Though we do not know exactly how the number of colonic crypts scales with body mass, estimates from human and mouse suggest that for every order of magnitude increase in body size, the number of crypts increase proportionally (see Methods). We used the same values as Calabrese and Shibata for all other parameters, which are listed in Table 2.

Parameter	Value	Definition
u	$3x10^{-6}$	Mutations/ongenic pathway/cell division
d	Age(days)/4	Divisions since birth (rate = 1 div./4 days)
k	6	Rate liming mutations required for cancer
N	8	Effective stem cells per crypt
m	$[1500 – 1.5x10^{10}]$	Crypts per colon

Table 2. Model parameters. These parameters used for the algebraic model to see how colorectal cancer incidence scales with body size. Parameter values were taken from (Calabrese and Shibata 2010). The mutation rate assumes that there are three genes (1Kb each) per pathway and a background mutation rate of 10^{-9} mutations per base pair per cell division.

If we use the blue whale as an example of an animal that is on the order of 1,000 times the size of a human, where m could equal 1.5×10^{10} crypts, this model predicts that all blue whales would have colorectal cancer by age 90 (Figure 7A). More specifically, when we solve the equation for years zero through 90 we find over 50% of whales would have colorectal cancer by age 50 and all would have colorectal cancer by age 80 (Figure 7C). The estimate for an animal 1,000 times smaller than a human (e.g. a mouse) is barely above zero even after 90 years. In reality, a mouse only lives a maximum of 4 years (de Magalhães and Costa 2009), so based on this equation they should never get colorectal cancer. The chance of an individual person getting colorectal cancer by age 90 is about 2.5% according to this model and 5.3% as reported by the American Cancer Society (ACS 2013). It is implausible that 100% of blue whales actually get colorectal cancer by age 80. Though we do not know how often blue whales are getting colorectal cancer, they have been reported to occasionally have other cancers (Martineau, Lemberger et al. 2002, Newman and Smith 2006) and can live for over 100 years (de Magalhães and Costa 2009). The poor fit of this model suggests that there is something

fundamentally different in the initiation and progression of cancer in large, long-lived animals compared to humans.

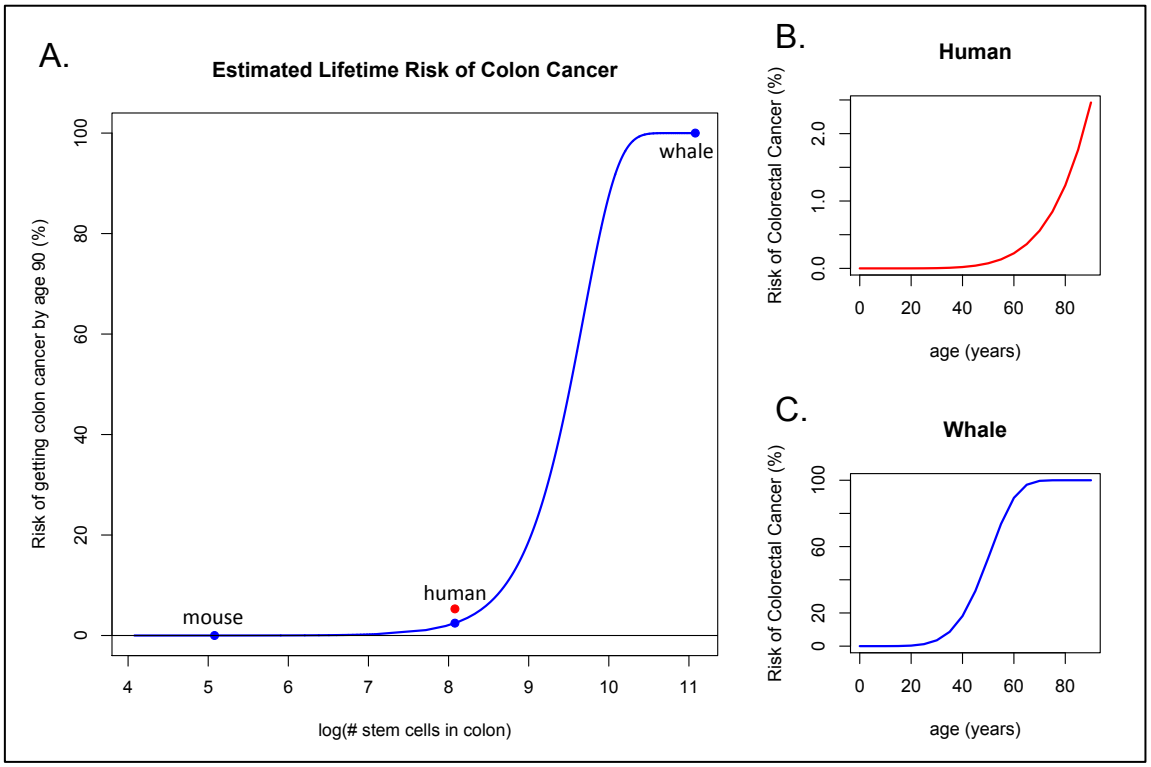


Figure 7. Estimated risk of colon cancer relative to body size. The probability was calculated using the algebraic model with the parameters listed in Table 2 (Calabrese and Shibata 2010) (A). Blue dots for mouse, human and whale indicate the estimated risk of colon cancer occurring within 90 years of life given the approximate number of cells in a human colon, 1,000 times fewer cells to represent the mouse, and 1,000 times more cells to represent the whale. The red dot indicates the lifetime risk of colon cancer according to the American Cancer Society which is about 5.3% for men and women averaged together (ACS 2013). The estimated age-incidence of cancer for human and whale, given this model, is shown in plots B and C respectively. This figure is adapted from the publication (Caulin and Maley 2011) with permission from Elsevier.

Next we investigated the set of parameter values that would allow the estimated age incidence of colorectal cancer in large animals to be similar to that of humans, which would match the empirical observation of Peto's paradox (Chapter 2). We tested 10,000

mutation rates ranging from $3.0x10^{-8}$ to $3.0x10^{-5}$ and solved for the value that minimized the difference between the estimates for human risk over 90 years and the calculated values for other species, given the number of colonic crypts. This analysis demonstrates that mere 3.2-fold decrease in mutation rate can account for a 1,000-fold increase in body size (Figure 8). The somatic mutation rates for an elephant and whale would need to be $4.6x10^{-10}$ and $3.13x10^{-10}$ respectively, in order for them to each have the same age incidence of colon cancer as humans (Figure 7B).

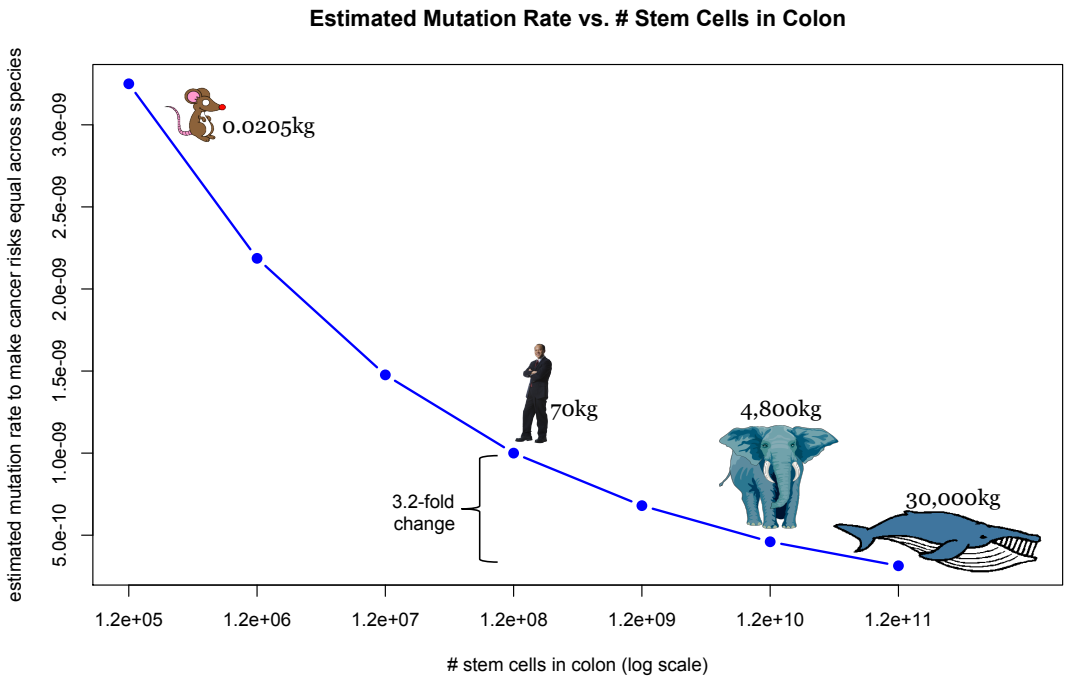


Figure 8. Estimated somatic mutation rates scaling with size. Mutation rate estimates show that a 3.2-fold decrease enables an animal that is 1000X larger than a human to have the same cancer risk. The mutation rates shown in the plot resulted in cancer risk predictions for the given number of cells that best matched the estimates for human (i.e. $1.2x10^{8}$ colonic stem cells).

Additionally we tested if altering the number of hits required for carcinogenesis (k) could allow cancer rates to be approximately equal across many orders of magnitude in size. Keeping all other parameters consistent with the values listed in Table 2, we varied k to range from 6-10. With 10 required hits, an animal 1000X larger than a human would have less than a 0.002% chance of getting cancer by age 90. However, just two extra hits (i.e. k=8) for an animal this size, gives the closet match to the human incidence curve (where k=6), and is slightly below with a lifetime risk of only 1.5%.

Another hypothesis that has been proposed to explain Peto's paradox involves changing the dynamics, or population size, of the dividing stem cells in structures such as crypts. With this model, we find that even if each crypt contained only one stem cell, a whale would still be predicted to have a lifetime colorectal cancer risk of 96%, so this is an unlikely solution to the paradox. However, changing the stem cell division rate from once every 4 days to once every 13 days for an animal with one thousand times more crypts than a human reduces the lifetime cancer risk to 2.2% and the age incidence line closely matches that of human.

Though our modeling efforts here are simplistic, they are still informative. We can likely rule out the possibility of large, long-lived organisms having fewer stem cells per crypt to explain Peto's paradox; however changing the division rate, mutation rate and number of required hits for carcinogenesis all seem feasible and the estimated values are in a normal biological range.

MODEL 2: WRIGHT-FISHER MODEL OF CANCER INCIDENCE

Our previous implementation of the algebraic model (Calabrese and Shibata 2010) of colorectal cancer ignores many dynamics of the cell populations. We incorporated slightly more realistic behaviors by implementing an adapted version of a previously published Wright-Fisher based model, which allows for cell lineage death (Beerenwinkel, Antal et al. 2007). We have simplified the model to maintain a constant population of size N, where N represents the entire population of crypt stem cells in the colon. This allows us to greatly reduce the computational complexity and more easily compare the results to the algebraic model (i.e. Model 1).

Using the same parameters that are in Table 2 and calculating colorectal cancer risk across orders of magnitude in stem cell number (i.e. body size), we find that the Wright-Fisher model provides a much lower estimate of lifetime risk. After 1,000 simulations of a human colon, the 90-year cancer risk is only 0.4% and for 1000-times as many stem cells, representing a whale colon, just over 25% of individuals get colon cancer (Figure 9). These lower values are expected when using the same input as in the algebraic model because the incorporation of random cell lineage death lowers the probability of a cell becoming cancerous since it not only has to accumulate all k mutations, but it also must avoid being eliminated from the population. However, 25% is still an extremely high rate when only considering one cancer type (i.e. colorectal cancer). In humans, the lifetime risk of most individual cancers are well below 10% with the exception of breast (12.4%) and prostate cancer (16.2%) (ACS 2013).

We also notice that the lifetime risk of colon cancer seems to level off around 25% for the largest species modeled (Figure 9). This inflection point is a consequence of the probability of losing a cell lineage becoming independent of population size when the population is sufficiently large in a Wright-Fisher model. As we explained in the beginning of this chapter, the probability that a given cell in generation *t* has no progeny in generation *t+1* is equal to $\left(1-\frac{1}{N}\right)^N$. As *N* increases we can make the following approximation:

$$\lim_{N\to\infty}\left(1-\frac{x}{N}\right)^N \sim e^x$$

Therefore, when *N* is sufficiently large, the probability of cell lineage death is independent of the population size and becomes a constant ($e^{-1} \approx 0.37$), which likely explains why cancer risk levels off of $N \geq 10^{10}$ with this model.

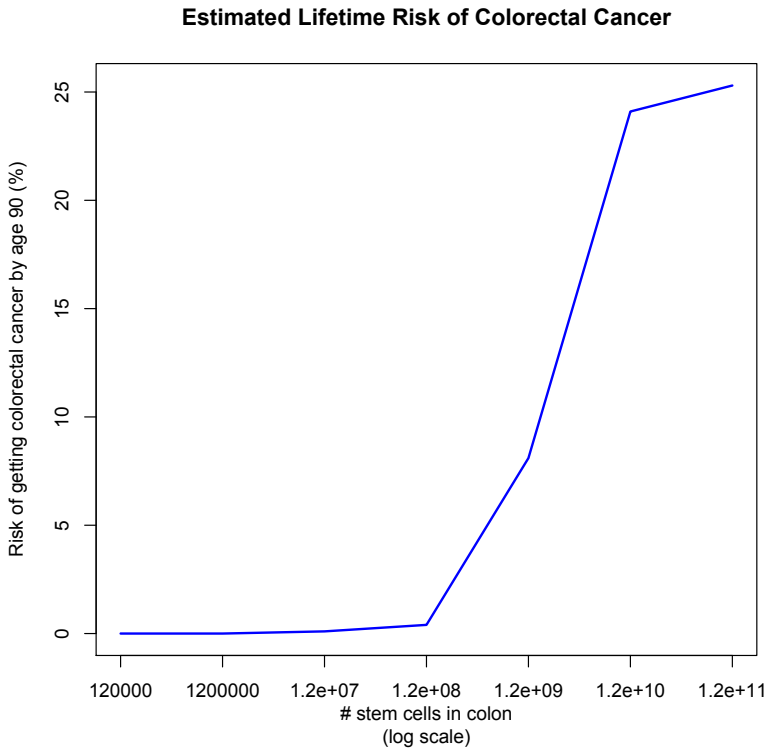


Figure 9. Updated estimates of colon cancer risk relative to size. Risk was calculated by running 1000 replicates of the Wright-Fisher based model three independent times for each stem cell value ($1.2x10^5 – 1.2x10^{11}$). According to this model and the parameters in Table 2, the human lifetime risk of colorectal cancer is 0.4% and 25.3% for an animal 1000 times as large.

As we had done with the previous model, we swept through different parameter values to find if the colorectal cancer incidence of whales could be lowered to match the estimate for humans. We show that just one additional required hit for colon cancer (i.e. k=7) can account for the risk due to the 1,000-fold increase in cell numbers. This one additional hit, which represents an extra pathway, decreases the lifetime risk of large animals, like whales, to 0.6% which closely matches the human estimate of 0.4% for k=6.

Decreasing the mutation rate for larger animals also greatly reduces their lifetime risk. Given $1.2x10^{11}$ crypt stem cells, a rate of $1.3x10^{-6}$ mutations per oncogenic pathway per division decreases the lifetime risk of cancer to the same as humans. This is only a 2.3 fold decrease from the value used to represent the human mutation rate in this model (compared to a 3.2-fold decrease we found in Model 1). Additionally, we show that this result can also be obtained by decreasing the cell division rate to once every 8.5 days. This results in a lifetime risk of 0.5% and a rate of one division every 9 days lowers this below the human estimate to 0.2%. However, decreasing the number of stem cells to be just one per crypt cannot sufficiently lower the risk to be comparable to humans, which is consistent with our results from the previous analysis.

Under this model, we estimate what parameter values allow the lifetime risk of colorectal cancer to be roughly equal between human and an animal 1000 times as large. The results are similar to what we obtained from the algebraic model; however the numerical changes are less extreme because of the effect that cell lineage death has on the random chance of a cell accumulating the necessary carcinogenic mutations.

METHODS

Justification for Assuming Colon Crypt count Scales with Body Mass

A human colon is on average 1.5 meters long and 6cm in diameter (Horton et al. 2000), which gives an approximate area of $3x10^3 cm^2$ (i.e diameter x π x length). The total number of crypts is estimated to be $1.5x10^7$ (Yatabe et al. 2001), so the crypt density is approximately 5,000 crypts per cm^2. A mouse, which is 3 orders of magnitude smaller

than a human, has roughly $6cm^2$ of colon (6cm long and 0.3cm in diameter) (Pickhardt et al. 2005). Using the same crypt density, we calculate there to be approximately $3x10^4$ crypts in a mouse colon, which is the expected 3 orders of magnitude change.

Algebraic Model

The algebraic model, which we have repurposed to explore solutions to Peto's paradox, was originally detailed in previous publications (Calabrese, Tavare et al. 2004, Calabrese and Shibata 2010). We use the same equation to calculate the risk of colorectal cancer given the age of the individual:

$$p = 1-(1-(1-(1-u)^d)^k)^{Nm}$$

where u is the mutation rate per gene per division, d is the number of stem cell divisions since birth, k is the number of rate limiting mutations required for cancer to occur, N is the number of effective stem cells per crypt and m is the number of crypts per colon (Calabrese and Shibata 2010). We wrote a script in C to run through the model using ranges for each parameter and the results were plotted in R.

Wright-Fisher Model

Dr. Trevor Graham wrote the code that was adapted for this analysis for his own implementation of the Wright-Fisher model of colorectal cancer. The model in our analysis maintained a constant population size with non-overlapping generations where each cell of the new generation choses a parent cell to inherit its mutant status from. This

occurs with equal probability (*1/N*) because we are not considering selective coefficients, as to make it more comparable to the algebraic model and avoid using parameters that lack good experimental measurements. Given a population of *N* cells, the probability of a configuration of cells with 0 to *k* mutations at a given time (*t+1*) can be expressed using the following multinomial distribution:

$$[N_0(t+1), \ldots, N_k(t+1)] \sim \frac{N(t)!}{N_0(t)! \ldots N_k(t)!} \prod_{j=0}^{k} \theta_j^{N_j(t)}$$

where $N_i(t)$ is the population size of cells at time *t* with *i* mutations and θ_j is the probability that a cell in generation *t+1* will have *j* mutations:

$$\theta_j = \sum_{i=0}^{j} \binom{d-i}{j-i} u^{j-i} (1-u)^{d-j} \omega_i x_i(t)$$

This has been formally detailed in the original publication (Beerenwinkel, Antal et al. 2007). In our implementation, each instance of the model represents one colon with *N* crypt stem cells. For each set of parameters, the model was run 1,000 times in order to estimate the frequency of cancer. We ran a minimum of three independent replicates of the 1,000 runs to make sure the number of cases reported to have cancer (i.e. contain *k* mutations) was consistent and we averaged across the replicates. R was used to visualize and plot the data.

DISCUSSION

These models are not intended to accurately represent the complexity of neoplastic progression; however we can still gain insight into what hypotheses may feasibly explain the observation of Peto's paradox. Interestingly, we find that the values that can resolve Peto's paradox by decreasing the lifetime cancer risk in large organisms fall within normal biological constraints. We were most surprised by the seemingly small changes in mutation rate that can account for a thousand-fold increase in body size resulting in cancer rates equivalent to humans. Estimates of the human somatic mutation rate span orders of magnitude and range from 10^{-11} to 10^{-9} mutations/base/division (Chu et al. 1988, Loeb 1991, Strauss 1992, Drake et al. 1998, Jones et al. 2008). One study that derives somatic mutation rates from specific loci across eukaryotes finds that the per base mutation rates for human and mouse are 5.0×10^{-11} and 1.8×10^{-10} respectively (Drake, Charlesworth et al. 1998). This 3.6-fold decrease in mutation rate in human cells compared to murine cells is close to the results of our modeling, which suggest that a 2 to 3-fold decrease in mutation rate can account for a 1000-fold increase in body size. This effective decrease in mutation rate may be accomplished by having better DNA repair, more efficient removal of mutated cells, or less endogenous damage as a result of a lower mass-specific basal metabolic rate (Caulin and Maley 2011).

We were also able to resolve Peto's paradox by increasing the number of rate-limiting hits required for transformation and by reducing the rate of stem cell divisions. Both models show that with just 1-2 additional hits, the risk of cancer can be greatly reduced in large animals. Therefore, we might anticipate finding redundant pathways or

additional tumor suppressor genes that add robustness to existing pathways in animals that have evolved this tumor suppression mechanism to better combat cancer.

To sufficiently decrease the lifetime risk of colon cancer in large animals like whales, we estimated that the division rate would be once every 8.5 days to once every 13 days, depending on the model. Crypt stem cell in mice divide once a day (Snippert et al. 2010); however human measurements are limited and are estimated to be at least once per week (Kang and Shibata 2013). Because we do not have accurate measurements for human, this result is more qualitative, but stresses that only small changes would need to occur in animals much larger than humans. One could investigate this by measuring the mitotic index of colonic crypts across species spanning orders of magnitude in size. We obtained samples of dolphin and whale colon; however the tissue was too degraded for us to get accurate estimates.

Though we were able to use simple models to gain insight into a complicated disease; there are many assumptions that go into these models that we must acknowledge when interpreting the results. These models assume that all mutations are evolutionarily neutral. That is, they provide no selective advantage to the clone, and so do not drive a clonal expansion. The model also assumes a constant population size and a constant mutation rate. Additionally, all k mutations necessary for cancer are required to occur in one single cell, which ignores the possibility of cell cooperation (Axelrod et al. 2006) and does not address clonal expansions, which would drastically alter the time to accumulate the mutations (Nowell 1976). The Wright-Fisher model (Model 2) was originally developed to model one single crypt as it progresses from a benign polyp to an invasive

tumor (Beerenwinkel, Antal et al. 2007). We have expanded the initial cell population to represent all stem cells in the colon; however, this ignores the compartmentalization structure provided by crypts since it allows a cell to provide more offspring in the next generation than just the population size of a single crypt. This simplification was made to drastically reduce the computational complexity, otherwise to model one colon would require running the model on a single crypt stem cell population millions-billions of times and then repeating this 1,000 times to estimate the cancer incidence. Our approach enabled us to run the analysis in a reasonable amount of time and allowed for more direct comparisons with the algebraic model (Model 1), which also did not consider the effects of the crypt structure.

We did not run these models to find exact numerical values for each parameter in a whale, as these will likely all vary depending on the exact model being considered. Rather, the goal of this analysis was to gain theoretical insight into the most realistic hypotheses to resolve Peto's paradox. We found that decreasing the mutation rate or division rate, or increasing the number of required hits can all sufficiently reduce the lifetime cancer risk in an animal orders of magnitude larger than a human; however decreasing the number of stem cells per crypts is not a likely solution. The necessary changes in the mutation rates and number of required hits are small and are well within biologically feasible ranges. These values could be the focus of future experiments designed to measure the somatic mutation rates and determine the number of pathways that must be mutated to transform cells across species that span a wide range of sizes.

CHAPTER 4: COPY NUMBER OF CANCER GENES IN MAMMALS

INTRODUCTION

Cancer-associated genes are generally divided into proto-oncogenes and tumor suppressor genes. Proto-oncogenes are defined as genes that increase the chance of progression to cancer when they are over-expressed or inappropriately activated (Adamson 1987). Tumor suppressor genes, on the other hand, are genes that increase the chance of progression to cancer when they are inactivated or deleted. Tumor suppressor genes often follow the 'two hit hypothesis' that requires both alleles to be mutated before causing a phenotypic change (Knudson 1971), though some are haploinsufficient such that inactivating a single allele is sufficient to cause a mutant phenotype.

Tumor suppressor genes are sometimes further divided into "caretakers" and "gatekeepers" (Kinzler and Vogelstein 1997). Caretakers help maintain genome integrity by preventing DNA damage and performing DNA repair. These functions evolved billions of years before multicellularity and are essential to all forms of life (Domazet-Loso and Tautz 2010). Gatekeepers control cell proliferation and signaling by enforcing checkpoints to ensure that cells at risk for neoplastic transformation do not continue to propagate. They do this by forcing cells to withdraw from the cell cycle via senescence or undergo programmed cell death (i.e. apoptosis) if the caretakers cannot repair them properly (Campisi 2005). Gatekeepers generally act in the interest of the whole organism

and often are not beneficial to a single cell, so it may not be surprising that many of them evolved with the emergence of multicellularity (Domazet-Loso and Tautz 2010).

As discussed in Chapter 1, the addition of tumor suppressor genes and the elimination of proto-oncogenes (or potentially oncogenic pathways) have been proposed as hypotheses to explain Peto's paradox. Additional tumor suppressors would provide robustness to the system and require a cell to accumulate a greater number of mutations in order to become malignant. Theoretically, gene duplication followed by persistence of a redundant function performed by each gene copy should not be evolutionarily stable because one gene can be altered without an immediate phenotypic consequence (Vavouri et al. 2008). However, the loss of redundancy in this case would lead to a cancer susceptible phenotype, and thus a reduction in fitness, allowing selection to maintain the redundant copies. It seems that functional redundancy is not simply a temporary result of gene duplication and can persist for many years after the duplication event. For example, the *MAP1* and *MAP2* genes, which are essential for cell proliferation, have both maintained duplicate functions in humans and yeast (*S. cerevisiae*), which diverged more than one billion years ago (Li and Chang 1995, Bernier et al. 2005). Though there may be no immediate phenotypic change if one copy of a duplicated tumor suppressor gene is deactivated, we predict that the selective pressure of increased cancer risk in large, long-lived animals is strong enough to stably maintain the functional redundancy and enhance cancer suppression.

The alternative to redundant tumor suppressor genes is the idea that the removal of proto-oncogenes or oncogenic pathways would minimize the number of vulnerabilities

within a cell and decrease the probability of an activating mutation occurring and promoting carcinogenesis. The 'null oncogene' hypothesis predicts that functional haploidy (meaning only one allele is active) at the loci of proto-oncogenes would reduce the risk of sporadic cancers (Davenport et al. 2002). Similarly, a decrease in cancer risk could also be achieved by removing some loci of proto-oncogenes so one could hypothesize that the number of proto-oncogenes in a genome might decrease as body size increases.

Because large bodies arose independently multiple times throughout evolution, we have no reason to believe that the copy number of a specific gene would scale with body size, as each lineage likely evolved different mechanisms to suppress cancer. Therefore, our initial hypothesis was that tumor suppressor gene families would expand and the total number of tumor suppressor genes (or possibly just caretakers or gatekeepers) would increase with body size. We also investigated whether or not the total number of proto-oncogenes decreased with body size. In subsequent analyses, we focused on specific tumor suppressor genes to discover if any were amplified in at least one genome of a large, long-lived organism.

EVOLUTION OF CANCER GENE FAMILIES IN MAMMALIAN GENOMES

If the evolution of large, long-lived animals involved the genomic amplification of tumor suppressor genes, these would appear as expanded gene families in those organisms. We developed a genome-wide BLAST search intended to find all genes within a gene family based on one representative. For example, we could accurately identify the extensively

studied *TP53* gene family (*TP53*, *TP63* and *TP73*) across species as our positive control by using *TP53* as the query gene. In order for a BLAST hit to be considered as an instance of the given gene family, we required that it pass several filters based on coverage, significance, function, and location (see Methods). We applied the BLAST search and filters to a highly curated set of 81 cancer genes to count the number of proto-oncogenes and tumor suppressor genes in eight mammalian genomes. We did not find a positive correlation between body mass and the number of genome hits for any of the cancer gene categories (one example shown in Figure 10).

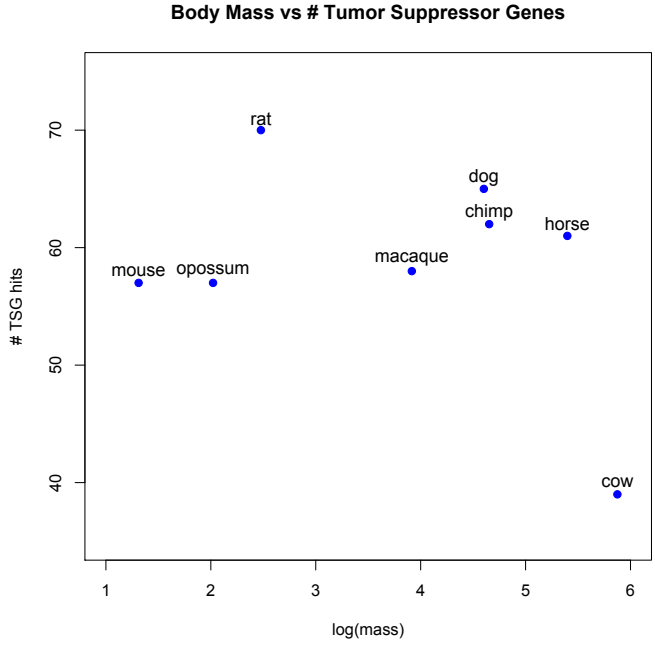


Figure 10. Number of tumor suppressor genes across mammals. The number of tumor suppressor genes does not increase with body mass. Based on our BLAST search we find no positive correlation between tumor suppressor genes as a whole, or gatekeepers and caretakers together with body mass. This was tested with a linear regression and is true on both the linear and log scale. The log (base 10) of the mass in grams is shown here to ease visualization of the range of masses.

Though we did not find a positive correlation between body mass and the number of TSGs as we had predicted, there is a week negative correlation with the number of gatekeeper genes ($r^2 = 0.66$, p-value = 0.015) and proto-oncogenes ($r^2 = 0.51$, p-value = 0.047). The relationship is also true for the combination of gatekeepers and caretakers; however, caretakers alone do not show any significant correlation with mass ($r^2 = 0.27$, p-value = 0.10). The negative association is driven solely by the lower counts found in cow and is completely abolished if the cow data point is removed from the analysis. Interestingly, we found a strong correlation between the number of proto-oncogene and gatekeeper hits, which seems independent of size ($r^2 =0.85$, p-value < 0.001) (Figure 11). We do not find this relationship between proto-oncogenes and caretakers ($r^2 = 0.13$, p-value = 0.36). Next we tested if any of the cancer gene categories correlated with lifespan or the product of mass and lifespan, since these features are highly correlated; however we found no significant relationships.

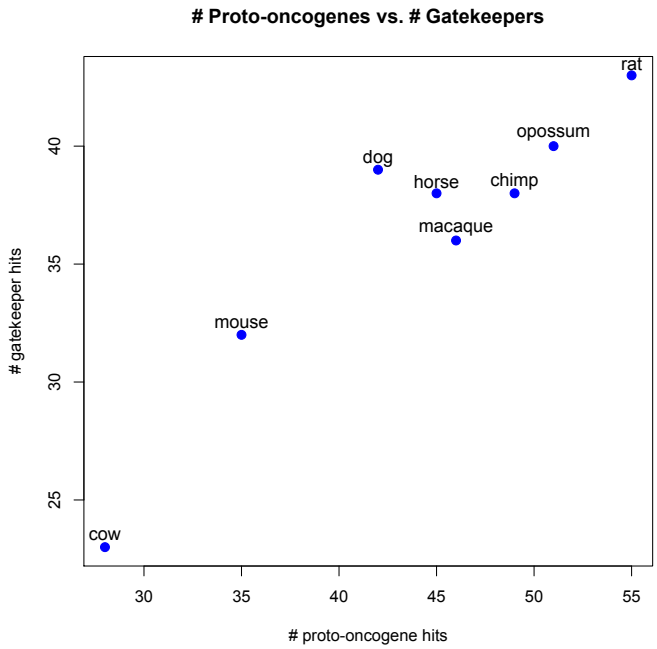


Figure 11. Correlation between proto-oncogenes and gatekeepers. There is a strong linear correlation between the number of proto-oncogenes and gatekeepers. Based on our BLAST search for cancer-gene families, the number of proto-oncogenes and gatekeepers found in a genome are highly correlated (r^2 =0.85, p-value < 0.001). Cow is the largest animal shown and has the lowest number of both gene types, though the rest of the data points are not in order of size.

COPY NUMBER OF TUMOR SUPPRESSOR GENES IN MAMMALS

Our BLAST analysis above is not sensitive enough to pick up small changes in individual gene copy numbers so we undertook a follow-up analysis to examine the copy number of specific tumor suppressor genes in mammals. We focused on increased copies of tumor suppressor genes since it is difficult to confirm a gene deletion in draft genomes due to possible incompleteness and misassemblies.

We used a comprehensive list of 830 human tumor suppressor genes (Higgins et al. 2007) and obtained the orthologous relationships in 36 non-human mammals from

Ensembl BioMart (version 72). Genes that were found to have a "one:many" relationship to the human tumor suppressor gene in at least on mammal were considered for downstream analysis. Our results revealed that 382 of the genes (46%) have at least one species with one or more additional orthologs to the human gene; though often these are listed in the database as "apparent orthologs" and are not high confidence calls. Only 11% of the genes (99) have 3 or more paralogs in at least one mammal and this decreases to a set of 36 genes (4.3%) when we filter on a minimum of 4 copies of a gene. To limit false positives due to the unknown certainty of low copy number increases, we focused on the instances of extreme gene amplification. We found that 19 tumor suppressor genes had five or more paralagous genes (i.e. at least 4 extra copies relative to the human genome) (Table 3). Some genes in the list (e.g. *IL6* and *CTGF*) are perhaps better known for oncogenic activity; however, they are included in the list of 830 genes because there are published reports of them demonstrating tumor suppressive behavior in certain tissues (Higgins, Claremont et al. 2007).

Our results show a number of interesting outliers with evidence of massive gene amplification (Table 3). The most extreme case is the *FBXO31* gene in the microbat (*Myotis lucifugus*) with 63 annotated copies. No other mammalian genome in the Ensembl database has more than one copy of this gene; however, the recent publication of the Brandt's bat (*Myotis brandtii*) genome reveals 57 copies of FBXO31 (Seim et al. 2013). This gene encodes an F-box protein that mediates the DNA damage response by promoting the degradation of *Cyclin D1* through polyubiquitination to induce cell cycle arrest in G1 (Santra et al. 2009). Though the microbat is only 10g, it can live up to 34 years (de Magalhães and Costa 2009) so one hypothesis is that these additional tumor

suppressors may decrease the cancer risk of the bat, that would otherwise be heightened by their increased longevity (Danilov et al. 2013).

Gene	Common Name	Scientific Name	Copy #
FBXO31	Microbat	*Myotis lucifugus*	63
TP53	African elephant	*Loxodonta africana*	12
IL6	Tree shrew	*Tupaia belangeri*	12
LCN2	Guinea pig	*Cavia porcellus*	12
CTGF	Lesser hedgehog tenrec	*Echinops telfairi*	9
ING4	Rock hyrax	*Procavia capensis*	9
ALOX15	Microbat	*Myotis lucifugus*	8
MAL	Horse	*Equus caballus*	8
MSMB	Opossum	*Monodelphis domestica*	8
	Guinea pig	*Cavia porcellus*	6
AKR1B10	Rat	*Rattus norvegicus*	7
LIF	Rock hyrax	*Procavia capensis*	7
	African elephant	*Loxodonta africana*	5
TCEB2	Rat	*Rattus norvegicus*	7
TNFRSF10A	Pig	*Sus scrofa*	7
TNFRSF10B	Pig	Sus scrofa	7
AKR1B1	Rat	*Rattus norvegicus*	6
SLIT2	Cat	*Felis catus*	6
CST5	Rat	*Rattus norvegicus*	5
IFNB1	Cow	*Bos taurus*	5
	Squirrel	*Ictidomys tridecemilineatus*	5
S100A11	Bushbaby	*Otolemus garnettii*	5

Table 3. Tumor suppressor genes amplified in non-human mammals. This list includes all tumor suppressor genes that we found to have at least four additional copies (i.e. 5 total copies) in mammalian genomes based on the "1:many" ortholog annotation provided by Ensembl.

The second highest gene copy number we came across was 12 which included *TP53, IL6* and *LCN2.* Because the focus of this work is on Peto's paradox, the 12 copies

of the canonical tumor suppressor gene *TP53* in the African elephant stood out as the most interesting. *TP53* is mutated in the majority of human cancers and plays a crucial role in multiple tumor suppressive pathways including apoptosis, senescence and DNA repair (Hollstein et al. 1991). Redundant copies of this gene could greatly reduce the risk of tumorigenesis and has been experimentally shown in mice (García-Cao, García-Cao et al. 2002). Chapter 5 will discuss the amplification of *TP53* in the African elephant in detail.

Additionally, the African elephant genome has 5 copies of *LIF* (leukemia inhibitory factor). *LIF* is a target of p53 and can induce cell differentiation in immune cells (Gearing et al. 1987). However, the closest sequenced relative to the African elephant, the hyrax (*Procavia capensis*), has 7 copies of *LIF*. When we looked at the mammals with less than 5 copies, we found that the lesser hedgehog tenrec (*Echinops telfairi*) also has 3 copies of the gene so we can assume that this amplification occurred before the speciation of these animals within Afrotheria and, though it may be biologically interesting, it is not likely an explanation to Peto's paradox.

The other species listed in Table 3 that are of interest include the horse (*Equus caballus*) and cow (*Bos taurus*). The horse draft genome (*EquCab2*) has 8 orthologs to the human tumor suppressor gene *MAL,* which are located in tandem on scaffold 15. The only other species in the database with any duplicate copies is the microbat with a total of two *MAL* loci. This gene is involved in T-cell differentiation (Alonso and Weissman 1987) and apical transport of membrane and secretory proteins (Cheong et al. 1999, Puertollano and Alonso 1999). Down regulation of this gene has been linked to multiple

epithelial cancers, including colon, cervical and esophageal (Mimori et al. 2003, Lind et al. 2008, Horne et al. 2009). The tumor suppressive properties of *MAL* have been verified in head and neck squamous cell carcinoma where the decrease of expression is associated with tumorigenesis and the exogenous expression of *MAL* decreased cell proliferation and increased apoptosis (Cao et al. 2010).

The final gene from our analysis with more than four copies in a large organism is *IFNB1* found in the cow. This gene belongs to the class of interferon genes known for their role in triggering the immune response to eradicate pathogens and tumor cells (Siegal et al. 1999, Takaoka et al. 2003). However, we also see the same number of redundant copes (5) in the squirrel genome and 2 copies (i.e. 1 extra copy) in the guinea pig, horse and hyrax genomes, which makes it less likely to be directly involved with enhanced tumor suppression in large, long-lived animals.

METHODS

BLAST Analysis for Gene Family Expansions

Curation of the cancer gene list was performed by Dr. Li-San Wang. We retrieved protein sequences of more than 300 genes from the Cancer Genome Anatomy Project (CGAP) website (Riggins and Strausberg 2001). We focused on genes with either oncogene (22 genes) or tumor suppressor (59 genes) classifications by CGAP. Other genes were classified as partners of fusion genes by CGAP, and were excluded from our analysis. We further divided the tumor suppressor genes into two groups: caretakers (CT; 28

genes) if the gene had gene ontology annotations suggesting their functionality in DNA damage repair; otherwise genes were classified as gatekeepers (GK; 31 genes). We used the NCBI gene ontology annotation for human, and checked for each gene if it is associated with a gene ontology term (or a descendant of such term in the gene ontology hierarchy) having "DNA damage" or "DNA repair" in its description.

Genomes from the NCBI RefSeq database were used as BLAST databases against the 81 human cancer related query genes in order to count the number of total hits in each genome. We limited the analysis to fully sequenced mammals: cow, chimp, dog, horse, macaque, mouse, opossum and rat. For a BLAST hit to count as an independent instance of that gene in a given genome it had to meet our criteria of coverage, significance, location, reciprocity and functionality. First, the union of all hits to that sequence in the subject's genome must cover at least 50% of the human query gene. Second, one of the BLAST hits in this region must have an e-value $\leq 10^{-5}$ and all other hits counting towards the 50% coverage must have e-values $\leq 10^{-3}$. Third, the BLAST hit must be greater than 1Mb away from any other determined location of the query gene in the given subject genome. The location of hits for each organism, based on these criteria, was used as input into the UCSC Genome Browser to retrieve the predicted protein sequences determined by the N-SCAN algorithm. These sequences were then used for a reciprocal BLAST back to human RefSeq protein sequences (release 37). In order for a region to count as a true hit in a non-human species, the predicted protein sequence must return a top hit in the human genome that is either the original human query gene that produced that hit, or a paralogous gene. Paralogs were defined by the Ensembl Genome Browser

(Release 56). N-SCAN was also used to determine the functionality of the genomic regions to exclude known pseudogenes and intergenic regions that were not predicted to be genes. These criteria were determined by comparison of our results to known p53 gene families (as reported by Ensembl release 56) as a positive control. The numbers of hits for each of the 81 individual genes were tallied as proto-oncogenes, caretakers and gatekeepers for each organism.

Body mass data (Smith 2003, de Magalhães and Costa 2009) and the evolutionary distance from humans was taken from the literature (Chen and Li 2001, Patterson et al. 2006, Bininda-Emonds et al. 2007, Gibbs et al. 2007, Murphy et al. 2007). We fit a linear regression model to the data using the statistical package *R* to determine the relationship between the number of each gene type (proto-oncogenes, caretakers, and gatekeepers) and the animals' body mass (representing the total number of cells in the organism). We tested this on both a log and linear scale.

Determining Copy Number of Tumor Suppressor Genes

A list of 830 tumor suppressor genes was downloaded from the Memorial Sloan Kettering CancerGenes database (Higgins, Claremont et al. 2007). This list includes all genes that have been associated with tumor suppressive behavior in at least one instance and have been assigned Gene Ontology terms related to these functions such as 'positive regulation of apoptosis' and 'negative regulation of cell proliferation'. Genes appear in this list regardless of whether or not they also have been reported to have oncogenic.

We obtained the orthologous relationships for 36 non-human mammals from Ensembl BioMart (version 72): alpaca, armadillo, bushbaby, cat, chimpanzee, common shrew, cow, dog, dolphin, African elephant, ferret, gibbon, gorilla, guinea pig, hedgehog, horse, kangaroo rat, letter hedgehog tenrec, macaque, marmoset, megabat, microbat, mouse, mouse lemur, opossum, orangutan, panda, pig, rabbit, rat, rock hyrax, sloth, squirrel, tarsier, Tasmanian devil, and tree shrew.

Genes that were found to have a "one:many" relationship, as annotated by Ensembl, to the human tumor suppressor gene in at least on mammal were considered for downstream analysis. The top genes were filtered based on the maximum number of times they occurred in any one species. All genes in Table 3 occurred at least 5 times in the species indicated.

DISCUSSION

We have analyzed the overall number of cancer-associated genes (divided into proto-oncogenes, gatekeepers and caretakers) in addition to a more detailed study of the copy number of individual cancer genes across 36 mammals. Our data does not support our initial hypothesis that the total number of tumor suppressor genes would increase proportional to body mass. Instead we see a trend in the opposite direction, where larger animals have fewer proto-oncogenes and gatekeepers. The association is eliminated if the cow is removed from the analysis. This may indicate that the correlation is an artifact of the cow draft assembly, though it is possible that the genome truly has fewer tumor

suppressors and proto-oncogenes compared to the other species. The correlation between the number of gatekeeper and proto-oncogenes suggests that there has been selection to balance the risk of the addition of an oncogene with the addition of a gatekeeper gene. This could also be evidence of the elimination of potentially oncogeneic pathways as discussed in Chapter 1 (Figure 2C).

A major caveat in this study is the difficulty in verifying a true gene deletion in a draft genome in the presence of incomplete assemblies, misassemblies, and inaccurate annotations. There may also be missing cancer genes in non-human species with little homology to the human gene sequences. However, we had added the time since the most recent common ancestor with human to our linear model and this did not change any results. Human tumor suppressor genes were used for this analysis, but in doing so we made the assumption that they perform the same function in the other species, which have not been experimentally verified. Additionally, we limited ourselves to these known tumor suppressor genes, but there may be additional genes acting as tumor suppressors in other species that would have been missed, in addition to possible flaws in our filtering criteria that could cause some genes to be missed. As an example, we set requirement that in order for two hits to be considered as separate instances of a query gene, they had to be at least 1Mb apart; however, if a gene were duplicated in a tandem repeat, we would likely only count as one copy.

Despite these limitations, we find a few genes that have been dramatically amplified in specific mammalian genomes, the most interesting of which is the discovery of 12 *TP53* copies in the genome of the African elephant genome. Another potentially

interesting gene in terms of Peto's paradox *MAL* which is found to have 8 copies in the horse genome and 2 in microbat. This could be an example of convergent evolution where a large animal (horse) and a small, long-lived animal (microbat) that both evolved extra copies of the same gene to overcome their increased risk of cancer. Further analysis and experimentation would need to be performed in order to determine the function of these copies and whether or not they provide enhanced suppression of carcinogenesis.

We chose to pursue the *TP53* amplification in the elephant to verify that their presence in the draft genome is not due to sequencing errors or misassemblies, and search for evidence of functionality. This work will be detailed in Chapter 5. Analyses such as this one are able to identify potentially interesting genes that may have a role in enhanced tumor suppression of large and long-lived organisms, and as more genomes become available gene amplifications in other species of interest can be brought to light for further investigation.

CHAPTER 5: AMPLIFICATION OF *TP53* IN AFRICAN ELEPHANTS

INTRODUCTION

There is selection upon the life history of organisms to suppress cancer long enough to maximize successful reproduction (Kirkwood 2005, DeGregori 2011). Selection has led to more effective cancer suppression in humans compared to mice, and more effective cancer suppression in elephants and whales compared to humans. However, the question still remains, how do large, long-lived animals suppress cancer better than smaller animals with shorter lifespans? Our results discussed in Chapter 4 suggest one possible mechanism of elevated cancer suppression: amplification of *TP53*.

TP53 (encoding protein p53) is a crucial tumor suppressor gene mutated in the majority of human cancers (Hollstein, Sidransky et al. 1991). Sometimes called the "guardian of the genome" (Lane 1992), p53 might be better known as the "Achilles heel of the genome." Many critical signaling pathways require p53, including DNA repair, apoptosis and cellular senescence (reviewed by (Ko and Prives 1996)). Inactivation of this protein can lead to suppression of apoptosis, increased proliferation, genomic instability, invasion and metastasis, four of the hallmarks of cancer (Hanahan and Weinberg 2011, Solomon et al. 2011). People lacking a functional copy of the *TP53* gene due to inherited germline mutations have Li-Fraumeni Syndrome. They have more than a 90% lifetime risk of cancer and are often diagnosed with multiple primary tumors (van Meerbeeck 1979, Gonzalez et al. 2009, Testa et al. 2013). Evolution appears to have left

an extreme vulnerability to cancer in our genomes. Additional functional copies of *TP53* would provide robustness that could help prevent carcinogenesis by increasing the number of mutations required to lose normal p53 function. In fact, mice genetically engineered to have extra copies of *TP53* show significant reductions in cancer (García-Cao, García-Cao et al. 2002).

VALIDATION OF *TP53* AMPLIFICATION IN AFRICAN ELEPHANTS

In the previous chapter, we found that the African elephant (*Loxodonta africana*) genome contained 12 *TP53* paralogs, according to the Ensembl database homology annotations. Further analysis, based on the alignments from the UCSC Genome Browser, revealed 20 copies of the tumor suppressor gene *TP53* in the draft genome *LoxAfr3*. We observe a large number of *TP53* copies reported in both Ensembl (12 protein coding copies and 1 pseudogene; release 72) and GenBank (1 protein coding copy and 19 pseudogenes) (Table 4). Apart from the differing numbers between the two databases, the automated gene annotations used by Ensembl predicted some abnormally short introns in the genes annotated as protein coding (e.g. 2 nucleotides), with only one copy having a gene structure comparable to that of *TP53* found in all other mammals (referred to hereafter as the *ancestral copy*). The other 19 copies, annotated as pseudogenes by GenBank, lack introns, which suggests they are a result of retrotransposition (referred to hereafter as *retrogenes* or *processed copies*).

We directly resequenced these multiple copies in order to verify their presence in the African elephant genome. Initially, PCR primers were uniquely designed for each copy (based on the 12 Ensembl *TP53* protein coding paralogs discussed in Chapter 4) and the products were sequenced. However, due to their high similarity, when the product was run on a gel and the target size bands were extracted, we found that each primer set amplified multiple retrogenes (Figure 12).

We overcame this challenge by cloning the *TP53* loci and sequencing 192 clones, which enabled us to confirm the presence of 18 retrogenes (GenBank accessions KF7185855-KF715872), all supported by multiple clones (Figure 13). Primers were chosen in conserved flanking sequences around the retrogenes to amplify approximately 2Kb fragments from each genomic location. Transformed colonies were picked, then the 2Kb fragment was amplified and linearized by PCR and purified for Sanger sequencing. We assembled full-length sequences of the inserted fragment for each clone that yielded quality sequence from all sequencing primers spanning the region. The cloned sequences, in addition to the published sequences, were aligned with MUSCLE (Edgar 2004) and this multiple alignment was used to construct a maximum likelihood phylogeny to determine the number of *TP53* copies that we were able to capture (Figure 13).

Eleven of the 18 sequenced retrogenes are similar but not identical to previous GenBank copies (Table 4). There was no evidence for eight of the published processed copies, which may be due to under-sampling of clones, misassembly in the published genome, or differences between individual elephants. An additional seven cloned

sequences have support from multiple clones but are not found in either database. There may also be additional copies of TP53 in the genome that were missed by our primers.

A.

B.

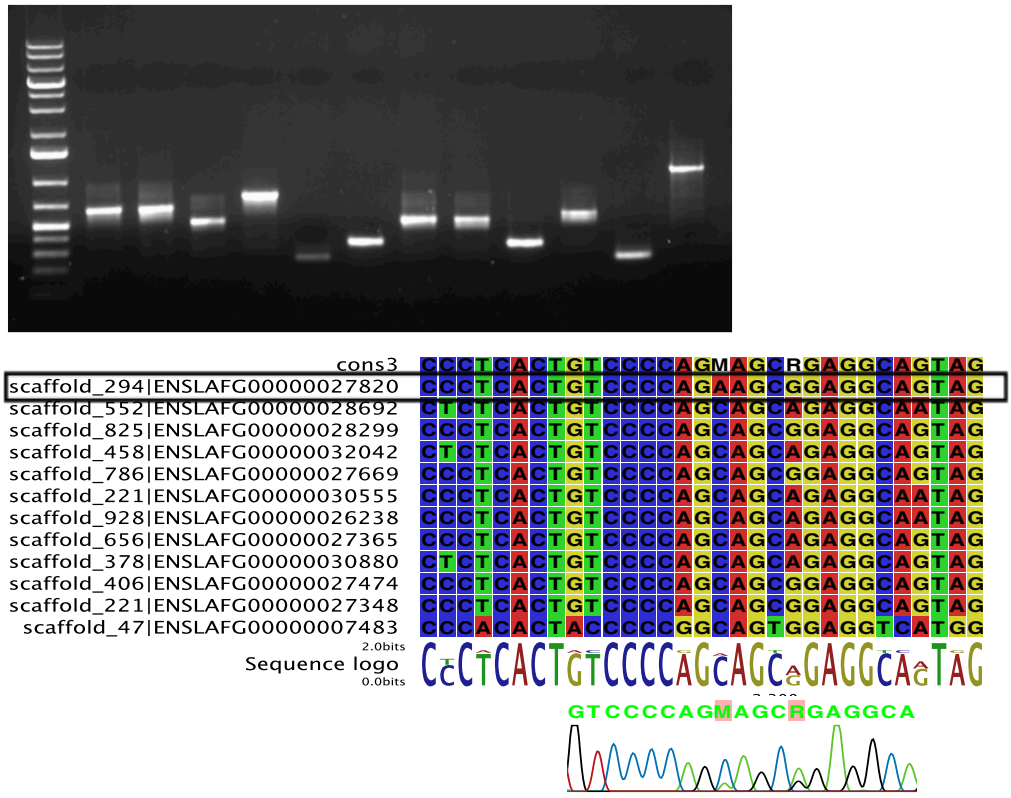


Figure 12. PCR products for 12 *TP53* copies in the African elephant. Primers for each of the 12 *TP53* copies annotated as protein coding genes by the Ensembl database (Table 4) were used in an attempt to amplify the different copies. PCR gave the expected sizes (A) and the bands were excised and sequenced. The sequence traces revealed a mix of retrogene copies present in each band. Here we show an example of the PCR product (top row of the alignment) from primers designed to capture the Ensembl gene ENSLAFG00000027820 (outlined by a black rectangle) aligned to the 12 Ensmebl sequences (B). The sequence trace below the alignment shows the heterogenous positions due to the mixture of loci in the sequencing reaction.

The percent identities of the pairwise alignments of each processed copy to the coding sequence (CDS) of the ancestral *TP53* copy range from 85-88%. The GenBank

sequence (gi:100663725) was used over the Ensembl prediction for the ancestral copy because it showed higher homology to other mammalian p53 sequences and had the expected exon structures. Within the processed copies, the percent identity ranges from 86-99% based on all pairwise alignments. The copies cluster into two groups of closely related paralogs: 6 in one cluster (Group A) and 12 in the other (Group B) (Figure 13). Between groups A and B the average percent identity is approximately 88% and within groups the percent identity is greater than 95% on average.

Additionally, we find evidence of conservation within the sequences flanking the retrogenes. The 574bp downstream of the suspected stop codon range from 83-86% conservation between the retrogene sequences and the 574bp downstream of the ancestral coding sequence. Pairwise comparisons of retrogenes show that the 260 bases 5' to the assumed start codon are not as highly conserved between the retrogenes and the ancestral copy and have percent identities ranging from 57-61%. However, the 14 bases immediately 5' of the start codon are conserved with 100% identity across all 18 retrogenes and the ancestral gene. The transcript of the ancestral *TP53* in the African elephant has yet to be sequenced and annotated so the boundaries for the 5' and 3' untranslated regions (UTRs) are currently unknown. Upon further analysis of the 5' and 3' sequences, we discovered that each copy is sandwiched between two mammalian interspersed repeats (MIRs), as annotated by RepBase, which may play an important role in how this gene has increased its copy number in the elephant genome.

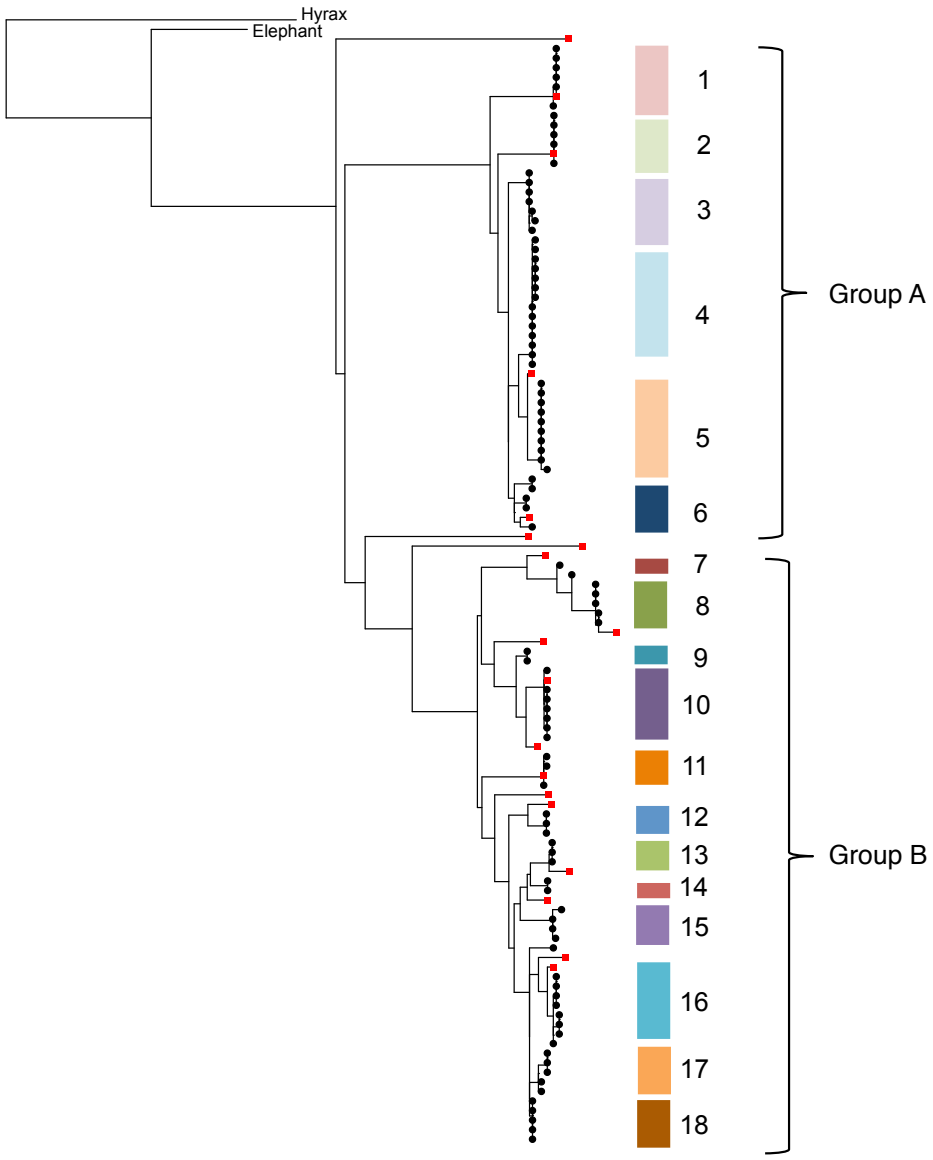


Figure 13. Phylogeny of TP53 clones in the African elephant. The African elephant genome has at least 19 copies of *TP53*. Capillary sequencing of retrogene clones reveals 18 distinct clusters of processed *TP53* copies (shown as colored blocks numbered 1-18) in a maximum likelihood phylogeny. The genes split into two main groups (labeled Group A and Group B). The sequenced clones are shown as black circles and published sequences from GenBank are shown as red squares. The branch labeled 'elephant' is the coding sequence of the ancestral *TP53* and 'hyrax' represents the coding sequences from the hyrax *TP53*.

We sought to determine when in evolutionary time the processed retrogene copies of *TP53* were introduced into the elephant. The hyrax (*Procavia* capensis) is the most closely related species to the African elephant with a sequenced genome and contains only one copy of *TP53* based on the draft genome *proCap1*. These lineages diverged between 54 and 65 million years ago (Eizirik et al. 2001, Kitazoe et al. 2007), placing a rough upper bound on the age of the introduction of the processed *TP53* sequences.

EVIDENCE OF TRANSCRIPTIONALLY ACTIVE RETROGENES

Next we tested if the extra processed copies of *TP53* could contribute to cancer resistance in elephants and thus help to explain the phenomenon of Peto's paradox. RNA was collected from treated and untreated PBMCs at various time points. The RNA samples were treated with DNase I and reverse transcribed with poly-T primers to create cDNA for downstream analyses. Primers were designed to distinguish the processed p53 copies from the ancestral sequence by spanning a region where Group A and B processed genes had different length deletions relative to each other. We performed PCR on the cDNA to look for presence of the retrogene transcripts. The primers allow for some degenerate binding and do not match all of the copies exactly, but if they were to bind and form products the ancestral copy, group A and group B of the processed genes would be 214, 201 and 185 base pairs long, respectively. The primers have highest identity to the retrogenes and after running a 1.5% agarose gel with the PCR products we see two bands, one at size 201bp and one at 185bp (Figure 14). The forward primer is located in the homologous region spanning exon two and three and the reverse primer is in exon three

which eliminates the possibility of the multiple products being from different isoforms of the ancestral *TP53* transcript. The bands were excised from a 10% polyacrylamide gel, which gave better separation when compared to agarose, and sequenced with Sanger sequencing to verify the copies (Figure 15). It was difficult to get a perfect cut separating the two bands with no contamination of one band to the other, so the sequence traces have some noise, likely from carry-over of the other PCR product. However, in instances where the sequencing is clean, it does appear that there are multiple retrogenes expressed from both group A and group B because we can pick up heterogeneous positions that correspond to the expected sequences of the various copies.

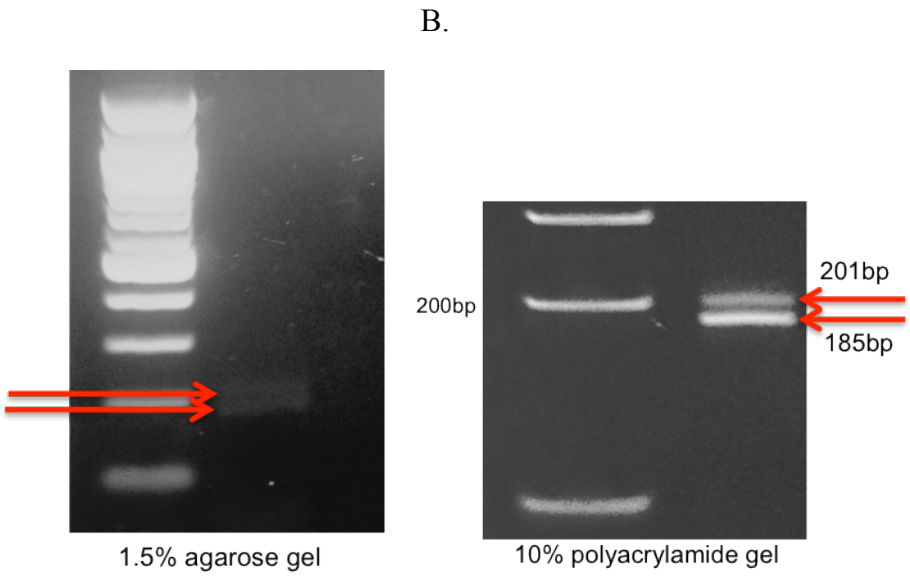


Figure 14. PCR products of TP53 transcripts. The p53 retrogenes are actively transcribed. RNA isolated from irradiated and non-irradiated elephant PBMCs was treated with DNase I, reverse transcribed with poly-T primers and PCR amplified with primers to distinguish the two groups of *TP53* retrogenes (groups shown in Figure 13) from the ancestral transcript. Shown here is the PCR product from cDNA prepared from the RNA sample at 5hr after treatment with 2Gy IR run on 1.5% agarose (A) and 10% polyacrylamide (B) for better separation and clarity of the bands.

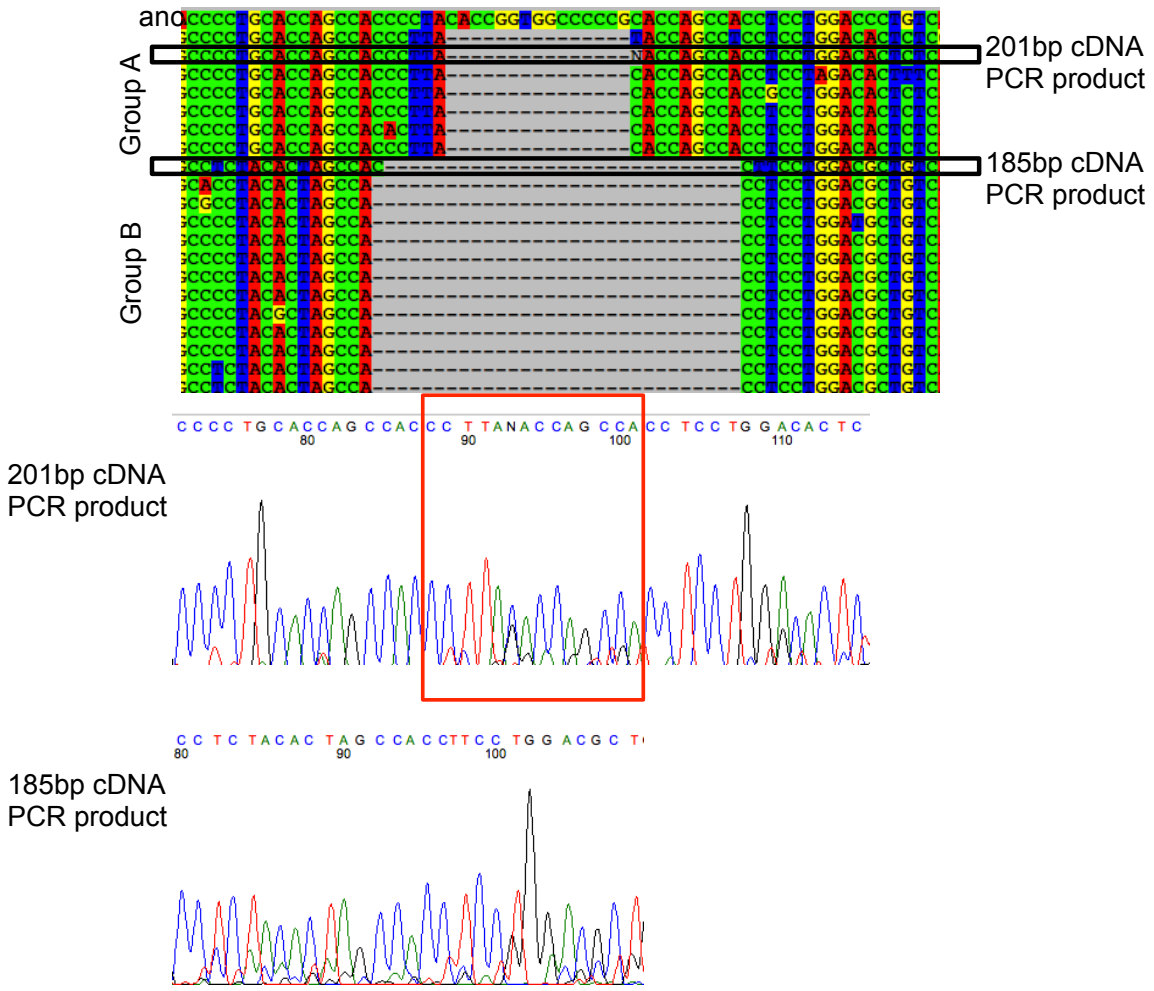


Figure 15. Capillary sequencing of excised gel bands from PCR of elephant cDNA. The above alignment is a screenshot from Seaview showing a segment of the ancestral *TP53* coding sequence, retrogene sequences from groups A and B and the aligned PCR product sequences (outlined by black rectangles) from the gel bands that were excised from the 10% polyacrylamide gel (Figure 14). This is evidence that retrogenes from both groups are actively transcribed. The sequences traces are shown below with the red box indicating the region that is deleted from the retrogenes in Group relative to those in Group A.

We were able to verify that at least a subset of the retrogenes were actively transcribed, so we ran qPCR in hopes of quantifying the total expression of *TP53*. Unfortunately, we were unable to design successful Taqman probes that could

differentiate between the retrogenes and the ancestral transcripts. We exposed both elephant and human peripheral blood mononuclear cells (PBMCs) to 2Gy gamma-irradiation to try and induce higher expression of *TP53*. qPCR was performed on RNA from the treated (2Gy IR) and untreated cells at time points 1hr, 5hr, 18hr and 24hr after exposure. *TP53* primers for the elephant samples were designed for the ancestral transcript only. We find no change of mRNA expression in human or elephant in irradiated cells compared to untreated cells. This is consistent with previous work showing that p53 is regulated at the protein level in response to DNA damage, thereby maintaining average levels of transcription (Giaccia and Kastan 1998). We performed a coupled *in vitro* transcription/translation assay to test if the retrogenes were able to produce proteins; however we did not find any evidence of successful translation.

APOPTOTIC RESPONSE TO GAMMA-IRRADIATION IN ELEPHANT CELLS

We hypothesized that if the extra copies of *TP53* are functional, elephant cells would either undergo more efficient DNA repair or have a higher rate of apoptosis compared to human cells when exposed to DNA damage. Better DNA repair would result in fewer somatic mutations while undergoing apoptosis in cells deemed 'too damaged to repair' prevents any mutations from that cell from being propagated through future generations. We collaborated with Dr. Joshua Schifman and Ashley Chan at the University of Utah to investigate radiation-induced apoptosis in elephants. We found that African elephant PBMCs apoptose at significantly elevated rates compared to human cells when exposed to γ-IR (Figure 16). PBMCs isolated from fresh African elephant blood, were exposed to

2Gy IR. Cells were stained with DAPI at different time points after exposure (1hr, 5hr and 24hr) and apoptotic cells were visualized and counted under the microscope. We analyzed the proportion of apoptotic cells at each time point in both treated and untreated samples. The difference of these two proportions (treated – untreated) was calculated along with the standard error of the difference. We tested if there was a significant difference in the amount of apoptosis in elephant when compared to human (see Methods for details). By 24 hours after 2Gy IR treatment, significantly more elephant cells had undergone apoptosis compared to human cells (Figure 16A).

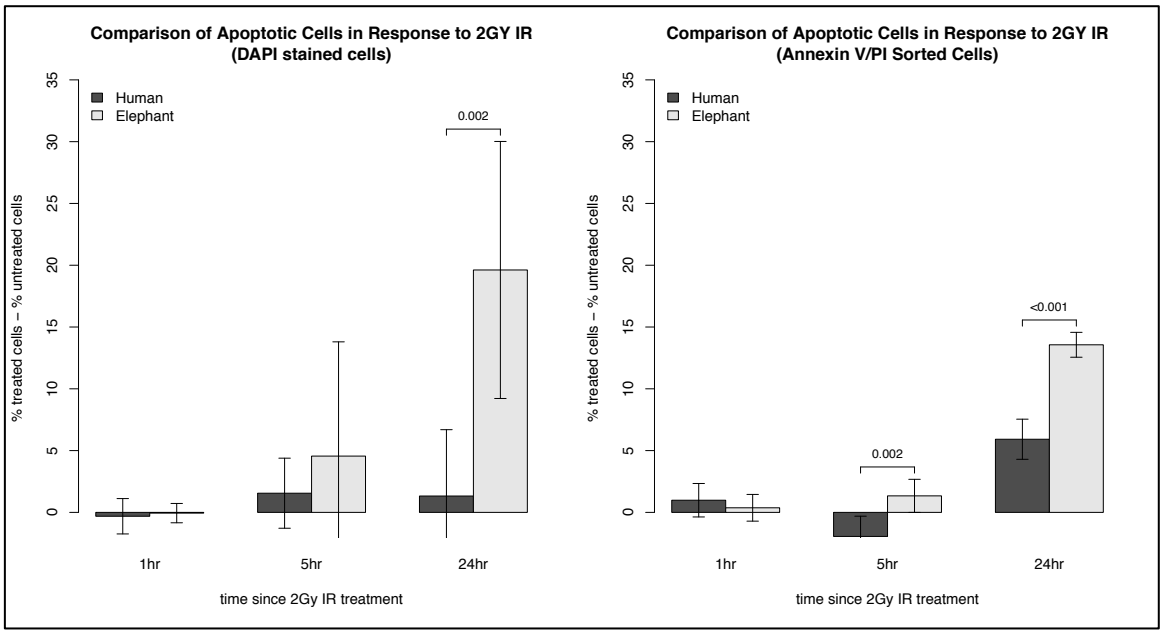


Figure 16. Elephant PBMCs are hypersensitive to gamma irradiation. Cells were stained with DAPI and apoptotic cells were counted at 1, 5 and 24 hours after treatment (A). The experiment was repeated using cell sorting based on Annexin V and PI staining (B) and supports the DAPI results. P-values less than 0.05 are shown above pairs of bars. The error bars represent the 95% confidence interval. Negative values are a result of more untreated cells undergoing apoptosis at that time point than the treated cells.

To verify the DAPI results, treated and untreated cells were stained with Annexin V and propidium iodide (PI) and sorted by flow cytometry at each time point. Cell sorting was able to verify that elephants have significantly more apoptotic cells in response to IR than the human samples (Figure 16B). The increased amount of observed apoptosis is not due to more DNA damage in elephant cells than human cells. We have confirmed this by counting pH2AX foci in treated and non-treated cells over 24 hours for both human and elephant PBMCs (Figure 17). These data suggest that the threshold for the amount of DNA damage that is tolerated by a cell for repair is lower in elephants than in humans.

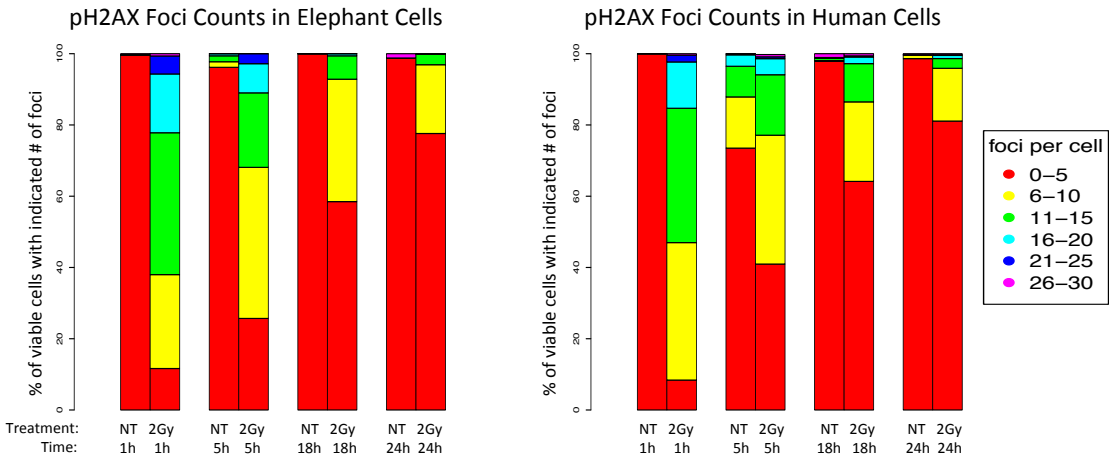


Figure 17. pH2AX foci counts in elephant and human PBMCs. PBMCs. There is no significant difference between the percentage of cells with pH2AX foci after 2Gy IR treatment in elephant (left) and human cells (right). NT indicated "no treatment" by irradiation and 2Gy indicates the intensity of irradiation.

To test whether this apoptotic response is specific to African elephants or is also observed in other species of elephants, we obtained fresh Asian elephant blood and repeated the IR experiments. Our results clearly show that Asian elephant PBMCs

undergo the same increased rate of apoptosis relative to human cells when exposed to 2Gy IR (Figure 18A). We also observe that the severity of the apoptotic response due to treatment decreases with age. The youngest Asian elephant (7 years) has approximately 30% apoptotic cells due to IR, while the older elephants (27-35) all are have less than 20% (Figure 18B). Our previous results that aggregated the different elephants are not driven by the extreme apoptosis in the young elephant because each individual elephant shows significantly increased apoptosis ($p < 0.05$) due to IR relative to the human samples.

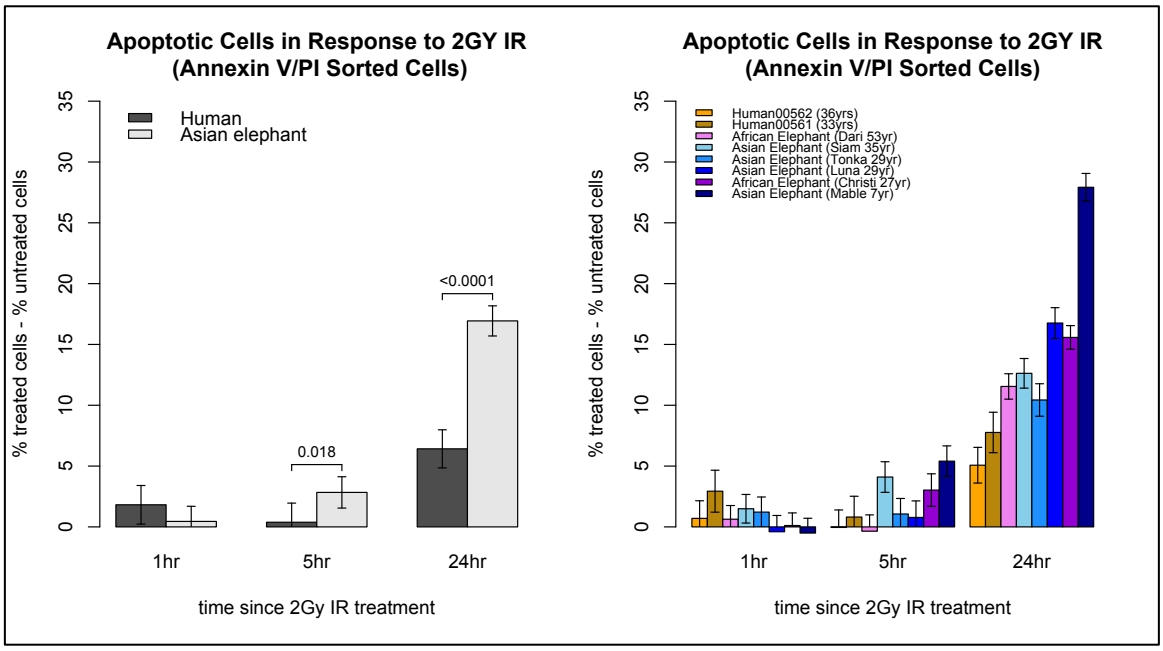


Figure 18. Apoptotic response in human, African elephant and Asian elephants. Asian elephant cells undergo significantly more apoptosis in response to 2Gy IR when compared to human cells (A). The experiment was done using cell sorting based on Annexin V and PI staining. P-values less than 0.05 are shown above pairs of bars (A). The error bars represent the 95% confidence intervals. Each biological replicate is shown for humans, African and Asian elephants (B). The apoptotic response appears to decrease with age. Ages are noted in the legend.

We performed RNAseq on the African elephant and human RNA collected from treated (2Gy γ-IR) and untreated cells at the five-hour time point. We mapped the reads to their respective draft genomes using Tophat (Kim et al. 2013) and calculated the transcript abundance of each gene with Cufflinks (Pollier et al. 2013). Cufflinks was used to calculate the normalized FPKM (fragments per kilobase of exon per million fragments mapped) to compare the treated and untreated expression levels. Cellular responses to γ-irradiation are largely driven by protein modifications, so we examined the expression of genes that are transcriptionally activated in p53-dependent apoptosis, cell cycle arrest and DNA repair (Soissi 1996, Tokino and Nakamura 2000) (Figure 19). We found that the human PBMCs show more drastic induction of genes compared to the African elephant; however this may be due to the fact that the transcriptome mapping is guided by gene annotations and in the elephant these are all computational predictions so there may be disagreements with the actual transcripts that are resulting in poor mapping. An increase in *BAX* expression is highly associated with p53-dependent apoptosis (Miyashita et al. 1994) and we see this in both human and elephant. There is little change in the other genes shown in Figure 19 in the elephant, suggesting that the apoptotic pathway is responding to the irradiation more strongly than the DNA repair and cell cycle arrest pathways.

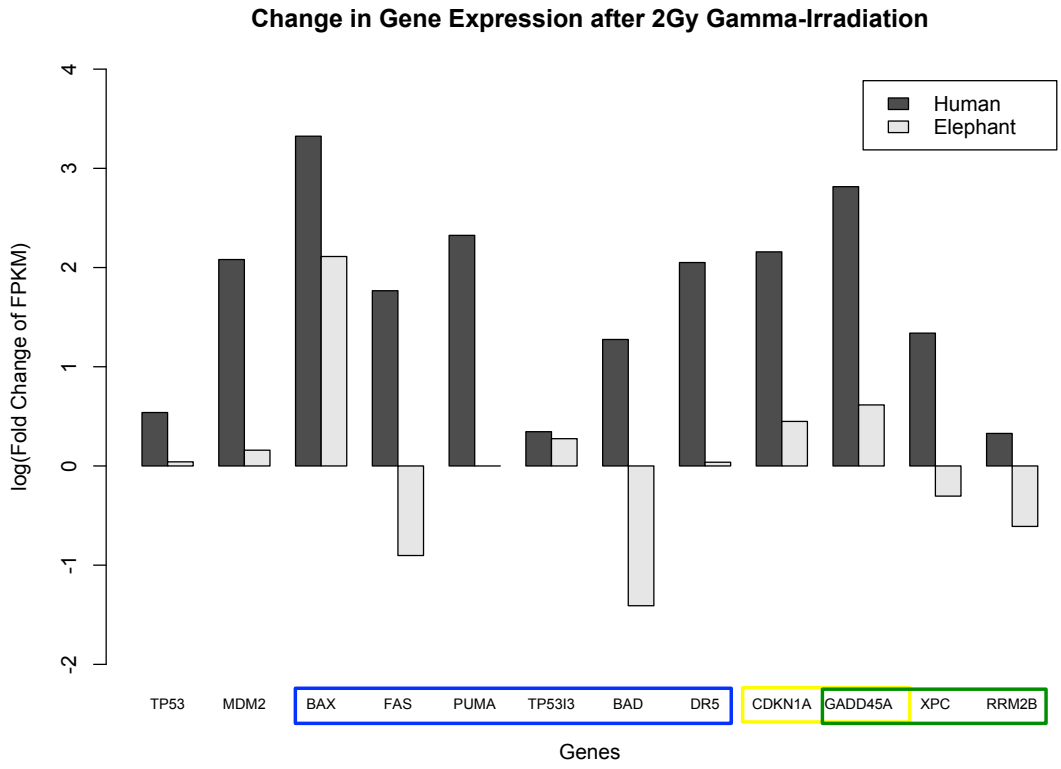


Figure 19. Induced Gene Expression after 2Gy Irradiation. The log-ratio of FPKM values (fragments per kilobase of exon per million fragments mapped) values of irradiated cells to untreated cells from RNAseq data for African elephant and human PBMCs were compared. Genes that are expressed during p53-dependent apoptosis (blue), cell cycle-arrest (yellow), and DNA-repair (green) are shown. *PUMA* is not annotated in the elephant genome.

Previous studies have shown that tissues with higher baseline expression of *TP53* mRNA are more sensitive to DNA damage (Komarova et al. 2000). We compared the human and elephant FPKM values for the baseline in untreated cells and our results suggest that elephants have slightly higher mRNA expression of *TP53* (8.15 FPKM vs. 5.61 FPKM). Technical and biological replicates would need to be performed with RNAseq in order to determine if this relationship is significant and may contribute to the heightened apoptotic response in elephants.

We also compared the expression of *MDM2*. The MDM2 protein is a negative regulator of p53 and is shown to increase in expression in response to DNA damage to ensure that p53 returns to the basal level in surviving cells (Perry 2004). The baseline FPKM values in untreated elephant and human PBMCs are 80.4 and 39.1, respectively. However, after exposure to 2Gy gamma-IR, the human expression jumps up to 165.4 and the elephant *MDM2* only increases to 89.8. Lower *MDM2* expression has been associated with increased sensitivity to radiation in human cells *in vitro* (Grunbaum et al. 2001), which suggests that the p53 protein in elephant is not being negatively regulated as strongly as in humans and may be acting more effectively to induce apoptosis.

MATERIALS AND METHODS

Sequence analysis

The African elephant genome assembly *LoxAfr3* was used for sequence analyses. *TP53* gene locations and sequences from the African elephant genome were obtained from the Ensembl database (release 72) and NCBI GenBank. From these two databases we find a total of 20 genomic positions that are homologous to *TP53*. The UCSC Genome Browser was used to view these 20 regions and verify that *TP53* transcripts of other species mapped to each of these locations. The start and stop of these locations was manually re-annotated based on these alignments to update some annotations that had truncated parts of the region homologous to the ancestral *TP53* coding sequence (Table 4). Multiple alignments were created with MUSCLE (Edgar 2004) and hand edited using Seaview. Percent identities between sequences were obtained from the MUSCLE output. PhyML

was used to create the maximum likelihood phylogeny for which gap sites were ignored. Flanking repeat sequences were annotated on the UCSC genome browser for the *LoxAfr3* genome and re-annotated with RepBase (Kohany et al. 2006) for the retrogenes we obtained from cloning and sequencing.

scaffold	start	stop	strand	Ensembl Gene ID	GenBank ID
175	436531	437673	+		100660069
208	307663	308805	-		100670203
217	57192	58321	-		100657221
221	32722	33852	+	ENSLAFG00000027348	100669451
221	320215	321342	+	ENSLAFG00000030555	100669732
281	127150	128270	+		100670118
294	64102	65230	+	ENSLAFG00000027820	100660838
342	119172	120298	-	ENSLAFG00000032258	100673852
378	23269	24394	+	ENSLAFG00000030880	100673452
406	137208	138342	-	ENSLAFG00000027474	100666240
458	14552	15678	+	ENSLAFG00000032042	100661323
47	11688313	11693871	-	ENSLAFG00000007483	100663725
498	44787	45912	+		100675551
552	13399	14524	-	ENSLAFG00000028692	100668616
627	40469	41597	+		100667946
656	10157	11282	-	ENSLAFG00000027365	100673935
76	9269289	9270442	+		100671320
786	1954	3080	+	ENSLAFG00000027669	100669552
825	4052	5178	+	ENSLAFG00000028299	100673857
928	6773	7899	+	ENSLAFG00000026238	100660953

Table 4. Genomic locations of 20 TP53 genes in the published LoxAfr3 genome. Genes with no support from resequencing of cloned loci are indicated in gray and black entries have high sequence identity with our sequenced clones. The ancestral copy containing introns is highlighted in red. ENSLAFG00000032258 is the one Ensembl pseudogene annotation.

Sample Collection

Whole blood samples from an individual male African elephant supplied by the Oakland Zoo were used for DNA analyses (cloning and re-sequencing). For the gamma-irradiation experiments and RNA collection, elephant whole blood samples were obtained from two African elephants at Utah's Hogle Zoo. Asian elephant blood was supplied by the Ringling Brothers Barnum and Bailey Center for Elephant Conservation. Human whole blood samples were obtained from healthy volunteers under IRB approved protocol at the University of Utah under the Cancer Genetics Study (CGS).

DNA Isolation

The DNA was purified with the DNeasy® Blood & Tissue Kit (QIAGEN) and concentrated by precipitating with 1/10 volume 3M sodium acetate (pH 5.2) and one volume of isopropanol followed by a 75% ethanol wash. The pellet was re-suspended in an appropriate volume of 1X TE buffer, based on the concentration determined by nanodrop, and stored at -20°C.

PCR and Capillary Sequencing of 12 Ensemble TP53 Genes

PCR primers were designed to be specific to each of the 12 copies of *TP53* in the African elephant genome. They were designed using Primer-BLAST and verified to be specific with the UCSC *in situ* PCR. Each 50uL PCR reaction contained PCR Buffer (Invitirogen), a 200uM concentration of each dNTP, 1.5mM concentration of $MgCl_2$,

0.2uM of each primer (Operon), 1 unit of Platinum® *Taq* DNA Polymerase (invitrogen), Q-Solution (Qiagen) and 50ng of template DNA. The following PCR program was used: 94°C for 2 min; 35 cycles of 94°C for 30 sec, 62°C for 30 sec, and 72°C for 1 min; and ending with 72°C for 5 min.

GeneID	**Primers**
ENSLAFG00000028692	5'-AACGAGTCAAAAGCCAGAAGCCACC-3'
	5'-GGGGGCAGTGCTTCACGACC-3'
ENSLAFG00000027365	5'-GCCACCATCCTGGGCACAGC-3'
	5'-GGTGGGGACAGTGCTGCACG-3'
ENSLAFG00000027820	5'-TGGGCTCTGGGGGCACCTTC-3'
	5'-CCACAGCTGCACTGGGCAGG-3'
ENSLAFG00000030880	5'-CCGAAGCCACCATCCTGGGC-3'
	5'-GCTCATAGGGCACCCCACGC-3'
ENSLAFG00000028299	5'-TGGGCTCTGGGGGCACCTTC-3'
	5'-TGCTGGGGACAGTGAGGGGG-3'
ENSLAFG00000007483	5'-AGGAAGTCGGGTGGGGAGCC-3'
	5'-GGCAGGGTGGGGACAGCAAC-3'
ENSLAFG00000027669	5'-TGCTGGGCTCTAGGGGCACC-3'
	5'-CCCACGGCTGCACTGGACAG-3'
ENSLAFG00000030555	5'-TAGCTGCTGGGCTCTGGGGAC-3'
	5'-ACAGCTGCACTGGACAGGCC-3'
ENSLAFG00000027348	5'-GCACCTGCTTTCTGGGCGTG-3'
	5'-AAGGGTGGCTGGTGCAGGGG-3'
ENSLAFG00000032042	5'-ATTTGCTTGGCCCCTGCCCTG-3'
	5'-GCCTCTGCTGCTGGGGACAG-3'
ENSLAFG00000027474	5'-GGCTCTGGGGGCACCTGCTT-3'
	5'-GCTGCTGGGGACAGTGAGGG-3'
ENSLAFG00000026238	5'-GGGAAGGGCTCTTCTGGGATGGTC-3'
	5'-AGGTGCTGGGCAGGGGTGTT-3'

Table 5. PCR primers for the 12 Ensemble elephant TP53 genes. These primers were used to amplify the 12 Ensembl *TP53* genes annotated to be protein coding in the African elephant. Their products are show in Figure 12.

The PCR reactions were run on a 1.5% agarose gel and visualized by UV light. The bands were extracted and purified using the PureLink™ Quick Gel Extraction Kit (Invitrogen). The products were sequenced at the Genomics Core Facility at UCSF and analyzed with the CLC Genomics Workbench 5 (CLC Bio).

Cloning and Capillary Sequencing of TP53 Processed Paralogs

Primers were designed to amplify the *TP53* processed copies simultaneously in the African elephant to be used in downstream cloning. 15 of the 19 published processed copies have sequences that perfectly match the primers and the others only differ by a couple of bases. All of the clones we sequenced were able to be captured with the one primer set (Forward 5'-GTCAGGTCACCTAGTTTCTGAATTG-3', Reverse 5'-GTCAATCCATCAACCAACAGG-3'). We also used a second reverse primer (5'-GTCAATCCATCAAAAAACAGG-3') with the same forward primer to try and match other copies more specifically, but the same loci were picked up with each set. Each 50uL PCR reaction contained 1X final concentration of PCR Buffer (Invitrogen), a 200uM concentration of each dNTP, 1.5mM concentration of MgCl2, 0.2uM of each primer (IDT), 1 unit of Platinum® *Taq* DNA Polymerase (Invitrogen), and ~50ng of template DNA. The following PCR program was used: 94°C for 2 min; 40 cycles of 94°C for 30 sec, 55°C for 30 sec, and 72°C for 2 min; and ending with 72°C for 5 min. Product sizes were verified by running the samples on a 1.5% agarose gel and visualizing bands with UV light.

The TOPO® TA Cloning® Kit for Sequencing (Invitrogen) was used to clone the PCR products into the PCR®4-TOPO® vector and transform chemically competent TOP10 E. coli cells. The kit protocol was followed for 3 cloning and transformation reactions: 2 replicates for the PCR products designed to capture the multiple copies of TP53 and one control transformation using the pUC19 plasmid. Each reaction was plated onto LB 100ug ampicillin agar at 20uL, 40uL and 100uL. Competent cells transformed with pUC19 were used as a positive transformation control while untransformed cells plated on ampicillin were used as a negative control.

To PCR the cloned product and verify the insert, 25uL reactions were prepared so that each reaction contained 2.5uL 10X PCR Buffer with MgCl2, 0.5uL dNTPs, 0.25uL T3 primer, 0.25uL T7 primer, 21.25uL water and 0.25uL Taq. A single isolated colony was scraped from a plate with a pipet tip and placed into a well containing the 25uL of PCR master mix. Two 96 well plates of colonies were prepared for a total of 192 reactions. The PCR program was run as follows: 94°C for 5 min; 30 cycles of 94°C for 30 sec, 55°C for 30 sec, and 72°C for 2.5 min followed by a 10 minute final extension at 72°C.

PCR reactions were diluted by adding 20uL water to 5uL of the PCR product and cleaned by adding 1uL SAP (1u/uL), 1 uL Exonuclease I (10u/uL) and 2uL SAP reaction buffer (USB® Products, Affymetrix, Inc). The samples (15pmol primer and ~150ng purified PCR product in a total volume of 6uL) were submitted to the Genome Core Facility at the University of California San Francisco for sequencing using ABI BigDye

v3.1 dye terminator sequencing chemistry and the new generation ABI PRISM 3730xl capillary DNA analyzer.

Four sequencing primers were used for each clone to fully cover the fragment with overlap: the T3 promoter 5'-AATTAACCCTCACTAAAGGG-3', the T7 promoter 5'-TAATACGACTCACTATAGGG-3', 5'-CCTGAGAAGCTGGTTCTGTCC-3', and 5'-CCAGACGTCAGCATATGATGGA-3'. Sequences traces were examined trimmed and assembled using CLC Genomics Workbench (CLC Bio).

Cell Culture

Ashley Chan performed the cell culture work in Dr. Joshua Schiffman's lab at the Huntsman Cancer Institute (University of Utah). Peripheral blood mononuclear cells (PBMCs) were isolated by Ficoll-Paque density-gradient and centrifugation followed by a red blood cell lysis. PBMCs were maintained in RPMI 1640 medium supplemented with 10% fetal bovine serum (FBS), 1% penicillin/streptomycin, and 2% L-glutamine. The cells were exposed to 2GY gamma-irradiation (γ-IR) in a RS-2000 X-ray Biological Research Irradiator followed by incubation at 37°C and 5% CO2 until given time point (1, 5, and 24 hours).

P53 Transcript Expression Assay

RNA was collected from the treated (2Gy) and untreated elephant and human cells at the 5-hour time point. The samples were stabilized with RNAlater and purified with the

Qiagen RNeasey Kit with on column DNase I treatment. Samples were reverse transcribed (RT) with the TaqMan® Reverse Transcription kit (Applied Biosystems) using poly-T and random hexamer oligos in separate reactions. The poly-T samples were used in downstream analyses.

Primers to distinguish the ancestral transcript from possible processed *TP53* transcripts were designed using Primer3 and their specificity to these regions was verified using BLAST against all predicted transcripts in the African elephant genome. The ancestral *TP53* cDNA was amplified using forward and reverse primers (5'-CCTCCTGGACCCTGTCATCTT-3' and 5'-AAGCCCAGACGGAAACCATA-3') and the processed p53 cDNA copies were amplified with forward and reverse primers (5'-CCTGAGAAGCTGGTTCTGTCC-3' and 5'-GCAGTAGGTCTTCTGGGAAGG-3'). Each 25uL PCR reaction contained a final concentration of 1X PCR Buffer (Invitrogen), a 200uM of each dNTP, 1.5mM of MgCl2, 0.2uM of each primer (Operon), 1 unit of Platinum® *Taq* DNA Polymerase (Invitrogen), and 1uL template cDNA directly from the RT reaction. The following PCR program was used: 94°C for 2 min; 35 cycles of 94°C for 30 sec, 55°C for 30 sec, and 72°C for 1 min; and ending with 72°C for 5 min. Product sizes were verified by running the samples on a 1.5% agarose gel and visualizing bands with UV light. Samples were also run on a 10% polyacrylamide gel to get better separation of the bands.

Bands were excised from the polyacrylamide gel and the DNA was purified with the Qiax II Gel Extraction Kit (Qiagen). The purified products were sequenced by Sanger

sequencing at the UCSF Genome Core Facility. Primers used for sequencing were the same as those for the PCR reaction.

Taqman probes were designed to compare the *TP53* expression in human and elephant after exposure to 2Gy IR. Primers and probes are listed in Table 6. We were unable to find Taqman primer/probe sets that passed the requirements for sequence composition and location as well as being able to uniquely target the retrogene transcripts. Instead we used a primer/probe set that was intended to only bind the ancestral *TP53* transcript. The qPCR was performed on the Applied Biosystems 7900HT real-time quantitative PCR machine at the Genome Analysis Core at the University of California San Francisco.

Primer/Probe	**Primer/probe Sequence**
LoxA GAPDH forward	5'-CCTGAGCTGAATGGGAAGCT-3'
LoxA GAPDH reverse	5'-TCAGATCCACCACTGACACGTT-3'
LoxA GAPDH probe	5'-ACT GGCATGGCCTTCCGTGTCC-3'
LoxA p53 Forward	5'-TGGGAACTCCTTCCTGAGAATC-3'
LoxA p53 Reverse	5'-TTCTGAGAGTAGCAGATCGTCCAT-3'
LoxA p53 probe	5'-TCCCCCACACTACCCCCGGC-3'
Human GAPDH forward	5'-ATTCCACCCATGGCAAATTC-3'
Human GAPDH reverse	5'-TGGGATTTCCATTGATGACAAG-3'
Human GAPDH probe	5'-ATGGCACCGTCAAGGCTGAGAACG-3'
Human TP53 forward	5'-CTGTCCCTTCCCAGAAAACCT-3'
Human TP53 reverse	5'-GCAGGGGAGTACGTGCAAG-3'
Human TP53 probe	5'-CCAGGGCAGCTACGGTTTCCGT-3'

Table 6. Taqman primers and probe sets used for qPCR.

RNA Library Preparation and Sequencing

RNA libraries were prepared and sequenced by the Genome Technology Center at the University of California Santa Cruz under the direction of Nader Pourmand. Two samples were run for both human and elephant (5hr treated (2Gy), and 5hr untreated (NT)). The RNA integrity (RNA Integrity Score >7) and quantity was determined on the Agilent 2100 Bioanalyzer following the manufacturer's recommendations. The total RNA (100ng) was treated by DNase using DNase mix from RecoverAll^{TM} Total Nucleic Acid Isolation kit (Applied Biosystems/Ambion) and subjected to cDNA synthesis with the Ovation RNA-Seq system V2 (Nugen).

RNA amplification was performed per the manufacturer's instructions and as described in detail in published literature (Tariq et al. 2011). Briefly, the total RNA was reverse transcribed to synthesize the first-strand cDNA by using a combination of random hexamers and poly-T chimeric primer. Double-stranded DNA (dsDNA) was generated by fragmentation of the mRNA template strand using RNA-dependent DNA polymerase. The dsDNA was purified using Agencourt RNAClean XP beads. The DNA was amplified linearly using a SPIA process in which RNase H degrades RNA in DNA/RNA heteroduplex at the 5'-end of the double-stranded cDNA, after which the SPIA primer binds to the cDNA and the polymerase starts replication at the 3'-end of the primer by displacement of the existing forward strand. Finally, random hexamers were used to amplify the second-strand cDNA linearly, as described previously (Tariq, Kim et al. 2011).

The double-stranded cDNA obtained after the Ovation V2 RNA-Seq system (0.5–

1 µg) was used for the library. The cDNAs were sheared down to 350-450bp using the manufacturer's protocol for the Covaris S2. A target insert size of 350-450bp was then size-selected using an automated electrophoretic DNA fractionation system, LabChip XT (Caliper Life Sciences). Paired-end sequencing libraries were prepared using Illumina's TruSeq DNA Sample Preparation Kit. Following library construction, samples were quantified using the Agilent Bioanalyzer per manufacturer's protocol. The libraries were sequenced using the Illumina HiSeq 2000 with sequencing paired-end read length at 2 x 100 bp. Reads were de-multiplexed using CASAVA (version 1.8.2).

RNAseq Analysis

Paired-end reads for each of the four samples (Human NT 5hr, Human 2Gy 5hr, Elephant NT 5hr, Elephant 2Gy 5hr), with an average of 105 million reads per sample, were filtered using fastq-mcf in the ea-utils package (Aronesty 2011). The adapter sequences were clipped from the ends and known Illumina artifacts were removed (a FASTA file of artifacts was provided by the Department of Energy Joint Genome Institute). The reads were then aligned to their respective genomes (human: *hg19*; elephant: *LoxAfr3*) with Tophat 2.0.6 (Kim, Pertea et al. 2013) and the FPKM (fragments per kilobase of exon per million fragments mapped) values were computed with Cufflinks 2.0.2 and analyzed for differential expression with *cuffdiff* (Pollier, Rombauts et al. 2013).

In vitro Translation

We followed the manufactures protocol for the TNT® T7 Quick Coupled Transcription/Translation System (Promega). Reactions were run with plasmid DNA isolated from transformed *E. coli* colonies and PCR products from the cloned 2Kb fragments in addition to a positive luciferase control (provided by the kit. We used Transcend® Biotin-tRNA as the free tRNA each reaction so that all successfully transcribed proteins were biotinylated and could be visualized by binding Streptavidin-Alkaline Phosphatase, as part of the Transcend Colorimetric Translation Detection System.

The reaction was subjected to SDS-PAGE. We followed the instructions for this from the coupled transcription/translation kit; however we found that denaturing the proteins at 70°C for ten minutes, instead of 100°C for two minutes, greatly reduced the background signal so we implemented this change. The gel was run in 1xSDS buffer for 1 hour at 100V and transferred to a PVDF (polyvinylidene difluoride) membrane by performing a semi-dry blot for 25 minutes at 15V. The membrane was blocked with TBST and incubated with Streptavidin-Alkaline Phosphatase for 60 minutes. Finally the proteins were visualized by incubating the membrane with Western Blue stabilizer.

DNA Repair Assay

Ashley Chan ran the DNA repair assays in Dr. Joshua Schiffman's lab at the Huntsman Cancer Institute. PBMCs were fixed with 2% PFA and adhered at $2x10^6$ cells per slide by

Cytocentrifuge. Immunofluorescence analysis was completed for detection of DNA repair by pH2AX foci staining with anti-gamma phospho-H2Ax (SER1399) (pH2AX) (Millipore Corp) (Wilson et al. 2011). Alexa Flour anti-rabbit IgG 594 and Alexa Flour anti-mouse IgG 488 (Invitrogen) were used to detect primary antibodies. Cell fields were visualized by confocal microscopy based on an Olympus BX41TF confocal imaging system. Images were captured at 100X with Picture Frame software (Optronics). Foci were manually counted at each time point and allocated to one of 6 categories (0-5, 6-10, 11-15, 16-20, 21-25, 25+).

Apoptosis Assays

Ashley Chan ran the apoptosis assays in Dr. Joshua Schiffman's lab at the Huntsman Cancer Institute. Flow cytometric analysis was done to investigate the percent of dead cells induced by IR. Apoptosis was detected by Annexin V staining. PBMCs were seeded in a 96 well plate at $2x10^5$ cells/well. Cell pellets were collected by centrifugation, washed with PBS and re-suspended with fluorescein isothiocyanate (FITC) Annexin V/propidium iodide (PI) (Apoptosis Detection Kit II, BD Biosciences). Cells were incubated in the dark for 15 minutes with the stain. Data were acquired with a Becton-Dickinson FACSCanto II laser cytometer (BD Biosciences). Readings were taken using nm excitation and band pass filters. Predictions for irradiated cell quadrant populations were based on x and y mean values previously determined by live statistics for non-irradiated cells (i.e. no treatment) at each time point with FlowJo software (TreeStar Inc.).

Apoptosis was also detected by immunofluorescence. PBMCs were fixed with 2% PFA and adhered at $2x10^6$ cells/slide by Cytocentrifuge. Cells were stained with DAPI and visualized for chromatin condensation and DNA fragmentation or stained with acridine orange/ethidium bromide (AO/EB) viability stain and visualized for late apoptotic cells. Apoptotic cells were manually counted based on cellular blebbing, chromatin condensation, and DNA fragmentation.

Statistical Analysis of Apoptosis

The same data analysis was performed for both the DAPI and Annexin V/PI experiments. For each sample at each time point, the proportion of apoptotic cells in non-treated samples were subtracted from the proportion in the corresponding treated sample (i.e. the same blood sample and time point). The standard error for the difference in these proportions was calculated with the following equation:

$$SE(p_{Gy} - p_{NT}) = \sqrt{\frac{p_{Gy}(1 - p_{Gy})}{n_{Gy}} + \frac{p_{NT}(1 - p_{NT})}{n_{NT}}}$$

where p_{Gy} and p_{NT} are the proportions of apoptotic cells in treated (indicated by Gy) and non-treated (indicated by NT) cells for a given time point. n_{Gy} and n_{NT} are the total number of cells (i.e. the sum of viable and apoptotic cells) counted for the sample (treated and untreated respectively). Three technical replicates were done for this experiment starting with fresh blood samples from the same elephants and humans each time. Each time point for an experiment is modeled with a normal distribution with mean $\boldsymbol{\mu}$ equal to

(p_{Gy} - p_{NT}) and variance $\boldsymbol{\sigma}^2$ equal to the square of the standard error (SE^2). We can combined the three experiments for a given individual elephant or human at each time point by averaging the three appropriate normal distributions using the following equation:

$$\frac{1}{3} \times [\mathrm{N}(\mu_1, \sigma_1^2) + \mathrm{N}(\mu_2, \sigma_2^2) + \mathrm{N}(\mu_3, \sigma_3^2)],$$

where the three normal distributions with means $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\mu}_3$ represent three independent technical replicates of the experiment. The DAPI experiment only had technical replicates, but in the Annexin V data there were 3 technical replicates for 2 elephants, 2 technical replicates for 2 human subjects and one replicate for another human subject. For the Annexin V data with biological replicates, the equation above is first used within the technical replicates of each individual and then applied again to combine the humans into one distribution $\mathrm{N}(\mu_h, \sigma_\mathrm{h}^2)$ and the elephants into another distribution $\mathrm{N}(\mu_\mathrm{e}, \sigma_\mathrm{e}^2)$ for each time point.

To test if the distribution for the composite elephant data is different than that of the composite human data, we performed a significance test for the difference in proportions. If σ_h^2 is the variance for the human data at a given time point and σ_e^2 is the variance for the elephant data at a given time point, then the standard error for the difference of the differences $(\mu_e - \mu_h)$ is equal to $\sqrt{\sigma_\mathrm{h}^2 + \sigma_\mathrm{e}^2}$. We then calculate the z-statistic where $z = \frac{(\mu_e - \mu_h)}{\sqrt{\sigma_\mathrm{h}^2 + \sigma_\mathrm{e}^2}}$ and find the corresponding p-value for a two-tailed test given a normal distribution.

DISCUSSION

Given the central role that *TP53* plays in response to cellular stresses including DNA damage, hypoxia, nucleotide deprivation and generally for cancer suppression (Lakin and Jackson 1999), it is surprising that humans have not evolved redundant copies. Early work in mice showed that there was a tradeoff between p53 activity and aging. Mice with constitutively active p53 were resistant to cancer but had accelerated aging (Tyner, Venkatachalam et al. 2002). However, follow-up work showed that adding extra copies of p53 to the mouse genome, under the endogenous promoter, generated mice that were cancer resistant and had a normal lifespan (Garcia-Cao et al. 2002). Mice in the wild usually die within a year from predation so there has been no selective pressure to suppress cancers that emerge over longer time intervals. In contrast, there has been strong selective pressure on elephants, which have on the order of 10,000 times more cells than a mouse, to suppress cancer long enough to bear and raise offspring over periods of decades. In addition, elephant herds with longer-lived matriarchs are more reproductively fit, providing further selective pressure to avoid cancer in aging elephants (McComb et al. 2011).

Our results show that *TP53* has gone through a massive expansion along the elephant lineage. It is likely that two retrotransposition events occurred, as indicated by the two main clusters forming groups A and B in the phylogeny of sequenced *TP53* loci (Figure 2) and distinguished by different short deletions and flanking sequences, followed by amplification via tandem duplications or retrotransposition of the retrogene transcripts. Additionally, our sequence analysis of the 5' and 3' regions surrounding the

retrogenes revealed that each copy is sandwiched between MIR (mammalian interspersed repeat) elements. MIR elements are SINEs (short interspersed elements) that belong to a non-autonomous class of retrotransposons. MIRs depend on the reverse transcriptase of LINEs (long interspersed elements) in order to complete reverse transcription and integrate into the genome in a new location. Interestingly, SINEs have been found to form composite transposons, which can mobilize the genomic sequence between two flanking repeats (Zelnick et al. 1987). This mechanism would disperse the *TP53* retrogenes around the genome, whereas local duplication would leave the genes clustered. Currently the retrogenes annotated by Genbank and Ensembl are all on small individual scaffolds and we re-sequenced very little upstream and downstream sequence of our clones, so there is no way to tell the genomic positions of these genes. Future sequencing of additional flanking sequence and mapping of each copy onto their respective chromosomes will reveal their physical proximity to each other and help to discover the mechanism behind the evolution of *TP53* in the African elephant.

Intronless retrotransposed genes, often referred to as processed pseudogenes, have previously been thought to be 'dead on arrival' (i.e. non-functional); however, there has been an increasing number of these genes found to be functional (Pink et al. 2011). Recent studies from the ENCODE project estimate that approximately 9% of all supposed pseudogenes in the human genome are actively transcribed and ~6% of processed pseudogenes are transcribed (Pei et al. 2012). A processed pseudogene of *TP53* found in rat has been shown to be transcriptionally active in response to heat shock in a histiocytoma cell line (Sreedhar 2010); however no further studies have been done to characterize the function of this gene. Unlike the pseudogene expressed in rat, which has

lost the DNA-binding domain (Sreedhar 2010), the elephant retrogenes have retained most of the sequence in the transactivation, DNA-binding and tetramerization domains, though not with 100% sequence identity. The largest deletion in the retrogenes is in the DNA-binding domain and is 15 bases long for retrogenes in Group A and 30 bases long for retrogenes in Group B.

We have shown that at least a subset of the retrogenes are actively transcribed in elephant PBMCs and that these cells undergo apoptosis at a significantly increased rate compared to human cells when exposed to 2Gy IR. This is evidence that the p53 pathway is more sensitive to DNA damage. We hypothesize that the retrogenes are increasing the cells sensitivity to DNA damage, and at lower thresholds of damage, these retrogenes are triggering p53-dependent apoptosis as opposed to DNA repair, which we find evidence of based on *BAX* expression in the RNAseq analysis. We also find a slightly increased baseline expression level of *TP53* in the elephant and we do not see a large increase in *MDM2* in response to IR, both of which can contribute to radiation sensitivity (Komarova, Christov et al. 2000, Grunbaum, Meye et al. 2001, Perry 2004). Interestingly, overexpression of *TP53* under an endogenous promoter in mice causes an increase in apoptosis when exposed to IR (García-Cao, García-Cao et al. 2002).

Apoptosis is an effective way to stop mutations from propagating to future cell generations. It has been shown that at critical points in embryonic development, murine cells are more likely to undergo apoptosis than repair DNA breaks (Heyer et al. 2000). Evolution may have enhanced this response in elephants to counteract the risk of cancer that should otherwise increase due to their large body size and long lifespan. We have

shown that this phenotype is found in both African and Asian elephants and that the response declines with age. A decline in apoptotic response has also been observed in aging murine T cells (Spaulding et al. 1997) and human sperm cells, despite an increase in DNA damage (Singh et al. 2003). We hypothesize that the higher apoptotic response found in younger individuals is due to an increased selective pressure to eliminate damaged cells and prevent carcinogenesis before sexual maturity. This increases the chance that the individual survives long enough and is healthy enough to successfully reproduce.

Each of the 18 retrogenes contains premature stop codons if they were to be translated with no splicing; however, they also all contain open reading frames (ORFs) longer than 100 amino acids. We did not find evidence of protein synthesis from the retrogenes with the *in vitro* coupled transcription/translation assay. This may be because our retrogenes included flanking sequence that interfered with proper transcription and does not prove that the *TP53* retrogenes do not function as proteins. It is possibly that these retrogenes are making truncated proteins, but we hypothesize that they are acting at the transcriptional level and are functional non-coding RNAs. Transcribed processed 'pseudogenes' have been found to occasionally act as microRNA decoys to increase the transcript level of the parent gene, as has been found with the tumor suppressor gene *PTEN* (Poliseno et al. 2010, Johnsson et al. 2013). Processed pseudogenes have also been shown to regulate the expression of their parent gene through translational interference by transcribing antisense RNA to bind to the parent mRNA or by being processed into siRNA (Pink, Wicks et al. 2011). One further possibility is that retrogenes in the African elephant genome target mutant *TP53* transcripts in the cell for degradation such that a

wild-type *TP53* allele would continue to form non-mutant homo-tetramers and provide the cell with fully functional p53 protein. Processed pseudogenes have also been associated with increased stability of their parent mRNA (Hirotsune et al. 2003), which may explain the altered response to IR in elephant cells.

The elephant's solution to Peto's Paradox is not necessarily the same solution that evolved in other large, long-lived organisms. Investigations into whales, naked mole rats, and other organisms with these extreme phenotypes should help illuminate the diversity of mechanisms that evolution discovered for suppressing cancer. Though we are not suggesting genetically engineering redundant copies of *TP53* into the human genome, there has been some evidence that small molecules may be able to restore function of mutant p53 (Rippin et al. 2002, Ventura et al. 2007) and others have been shown to stabilize the protein by disrupting the interaction with MDM2 (Issaeva et al. 2004, Vassilev et al. 2004, Shangary et al. 2008). The results of evolution in the elephant suggest that enhancing *TP53* function is a promising direction for cancer prevention in humans.

CHAPTER 6: *DE NOVO* ASSEMBLY OF THE HUMPBACK WHALE GENOME

INTRODUCTION

Few genomes of large, long-lived organisms have been sequenced, which greatly limits the ability of researchers to make progress on comparative studies of aging and age-related diseases, such as cancer. To address this information gap, we sequenced and *de novo* assembled the humpback whale (*Megaptera novaeangliae*) genome. Referred to as the 'drosophila of whales' among marine biologists (personal communication with Per Palsböll), the humpback whale has been studied in much greater detail than other whale species. There is a large community of scientists that track humpback whales, perform genetic testing, and decode their songs. Because of the amount information that has been accumulated about this species, we felt that providing a draft genome would not only further our interests in cancer research, but may provide the information needed to examine the genetic basis of many humpback whale phenotypes.

Whales are classified in the order Cetacea, which contains two suborders, Odontoceti, the toothed whales (e.g. dolphins), and Mysticeti, the baleen whales (e.g. the humpback). Mysticeti and Odontoceti share a most recent common ancestor approximately 20-34 million years ago (Murphy, Pringle et al. 2007, Jackson et al. 2009). Humpback whales have an observed maximum lifespan of 95 years and have an average adult weight of 30,000Kg (de Magalhães and Costa 2009), though they can grow to as much as 48,000Kg (Schmidly 1994). They are clearly a prime example of a large, long-

lived organism with which to study Peto's paradox. Very few cases of cancer have been reported in this species; however there are documented cases of benign neoplasms including a basal lipoma in the central nervous system and fibromas on the tongue and skin (Newman and Smith 2006). Humpback whales are closely related to species that are orders of magnitude smaller, such as the harbor porpoise (*Phocoena pohocoena*) (which we have plans to sequence with collaborators), which weighs approximately 52.5Kg (de Magalhães and Costa 2009). Close evolutionary relationships across magnitudes of body sizes make it more straightforward to determine genes involved with the evolution of large body size and long lifespan. Other research groups are beginning to sequence closely related cetaceans that span a range of sizes, which will enable many interesting comparative studies to be performed.

The individual female humpback whale that was used for this sequencing project is already well known in the marine biology community and by local whale-watchers in New England. Her name is Salt and she was the first whale in the world to be assigned a name (as opposed to a number). She was first spotted by Captain Aaron Avellar in 1975 in Massachusetts Bay who named her for the distinct white scar pattern on her dorsal fin (Knaub 2001). Salt has a history of helping researchers, as she was later seen that same year off the coast of the Dominican Republic which provided scientists with valuable information to understand the migration patterns of the North Atlantic humpback whales (NOAA 2006). She has been seen every year off the coast of Cape Cod for over 35 years, with the exception of one, and is estimated to be roughly 45 years old (Knaub 2012). Salt has mothered 12 calves and is a grandmother to 10 calves, all of which have

been named by Captain Avellar and his family, whom continue to monitor Salt and her family from year to year (Knaub 2012).

The humpback whale population has been reduced from a global population of more than 200,000 to near extinction as a result of unregulated whaling (Clapham et al. 1999). In recent years the population of humpbacks has been recovering and the current worldwide population is approximately 80,000, as reported by the International Whaling Commission (IWC 2013). Researchers have used genetic markers to look at variation within the existing populations. They expected to see low diversity due to the extreme population bottleneck, however the populations maintain higher levels of nucleotide variation among individuals than expected, which is thought to be a preservation of the past heterogeneity as opposed to a recent post-bottleneck explosion (Baker et al. 1993). A full reference genome would allow for more in depth genetic studies for comparative biology as well as to better understand the current population and perhaps aid in continued conservation efforts.

There are a number of other scientific communities that are eager for access to a baleen whale genome. The aging community would like to make use of this genome to gain insight into what genes may be responsible for the extended lifespans of whales. In order to transition successfully to marine life from their land-living ancestor, whales had to go through rapid adaptation and this evolutionary history remains encoded in their genomes, making this genome of great interest to evolutionary marine biologists seeking to understand this transition. We hope that publishing this genome will produce novel data that will be used in many fields of science from basic biology to biomedical research and could open new doors for cancer prevention.

THE COMPLEXITIES OF THE HUMPBACK WHALE GENOME

The *de novo* assembly of any mammalian genome from next-generation sequence data is a challenge; however the biology of the humpback whale genome makes this effort particularly difficult. Previous work on humpback whales revealed the karyotype diploid value (*2n*) is 44 chromosomes (Lambersten et al. 1988). The genome is assumed to be around three billion bases, based on measurements of DNA content in cells of toothed whales (Gregory 2013). Chromosomal hybridizations have shown that cetacean genomes are highly repetitive and studies have since focused on three major repeat sequences found in baleen whale genomes (Arnason et al. 1978, Arnason and Widegren 1989). About 15% of the genome is estimated to be composed of one tandemly repeated 1.7Kb sequence which is typically localized near the telomeric regions of the chromosomes (Arnason and Widegren 1989). A 72bp sequence within this repeat has dyad symmetry, and approximately 540 bases show similarity to the mammalian LINE-1 element (Kapitonov et al. 1998). This repeat is known as the common repeat and is found in all examined cetaceans from odontocetes (toothed whales) to mysticetes (baleen whales) (Arnason et al. 1984).

There is 422bp heavy (GC-rich) satellite found in baleen whales, which is also organized in tandem repeats (Arnason and Widegren 1989, Adegoke et al. 1993). Roughly one half of this repeat sequence is composed of the subrepeat TTAGGG, which is a common motif found in mammalian telomeres so it is not surprising that this repeat is distal to the common 1.7Kb component (Arnason and Widegren 1989). The third studied repeat is the light (AT-rich) satellite, which has not been sequenced but seems to be less conserved based on restriction hybridization patterns from eight species which show a

range of fragment lengths (Arnason and Widegren 1984). It localizes in the centromeric regions where it is thought that there may be fewer constraints on the evolution of repetitive sequences (Arnason and Widegren 1989).

Though next-generation sequencing technologies (NGS) have enabled us to sequence genomes more efficiently and for less money, resolving repeats with this data is an extreme challenge. Repeats that are longer than the read length, which includes these common cetacean repeats, can cause a fragmented assembly since they may not be able to be anchored to surrounding non-repetitive DNA (Treangen and Salzberg 2012). Additionally, repeats can be collapsed resulting in a shorter assembly and complex misassemblies (Phillippy et al. 2008, Pop and Salzberg 2008). The two best current methods to deal with repeats is by producing longer reads and having long inserts between mate-pair reads that can span repeat regions and help anchor them uniquely (Treangen and Salzberg 2012). We have pursued both strategies in our assembly of the humpback whale genome.

DE NOVO SEQUENCE ASSEMBLERS

Due to the decrease in sequencing costs, individual labs are now able to sequence a genome with next-generation and second-generation sequencing technologies; however, *de novo* assembly from these short reads (~100bp) remains a challenge. This has prompted many groups to develop assembly software, all of which have relative advantages and disadvantages. There are two main foundations upon which a non-greedy assembly algorithm can be built: de Bruijn graphs and overlap graphs (Miller et al. 2010)

(Figure 20). We have made use of both strategies, as well as a string-graph algorithm, which is a variation of the overlap graph, in order to generate the current draft assembly of the humpback whale.

Our first assembly of the humpback whale was done with the software ALLPATHS-LG, provided by the Broad Institute (Gnerre et al. 2011). This algorithm is based on a de Bruijn graph (Figure 20D), which are typically chosen for large data sets with millions of reads. Reads are broken down into *k-mers* and each string of *k* bases is a node. Nodes are connected by a directed edge if they overlap by *k-1* bases. De Bruijn graphs are known for their computational efficiency because they do not need to explicitly compute all pairwise overlaps, which is $O(n^2)$ for n reads. Instead, de Bruijn graph assemblers scale with the number of unique *k-mers* in the sequence data. Though this saves computational time, it requires a large amount of memory to store the graph, which has an upper-bound of $O(4^k)$. The default value for ALLPATHS-LG is *k=24* for the fragment libraries and *k=96* for jump libraries, which the developers encourage users not to change as many heuristics in their algorithm are highly sensitive to this value (Gnerre, Maccallum et al. 2011). Assembly of a mammalian genome therefore requires over 500GB of RAM, and the manual recommends a machine with 1TB of RAM.

A consequence of the de Bruijn graph approach is that the information contained in a full read is reduced to *k-mers* and therefore the information about *k-mer* adjacency is lost. Repeat regions are often collapsed and not easily resolved; however this method also eliminates spurious assemblies caused by repetitive sequences at ends of reads which can occur with overlap-based assembly methods (Miller, Koren et al. 2010). The major

limitation of ALLPATHS-LG and all de Bruijn graph assemblers is that they cannot accept a hybrid of sequence data as input, which is what ultimately convinced us to try a different assembler.

Overlap-based assemblers are more flexible then de Bruijn graphs and can use variable length reads as input (Myers et al. 2000, Zimin et al. 2013). Our most recent draft assembly of the humpback whale takes advantage of the new MaSuRCA 2.0.1 assembler (Zimin, Marcais et al. 2013), which is built around an adapted version of the Celera Assembler 6.1 (CA)(Myers, Sutton et al. 2000, Miller, Koren et al. 2010). CA is an overlap-layout-consensus (OLC) algorithm (Figure 20B) in which each node of the graph represents a full read and each directed edge *(u,v)* denotes an overlap between reads *u* and *v,* where the edge weight is equal to the maximum length suffix of read *u* that is a prefix of read *v* (i.e. the size of the overlap between read ends). In addition to increased flexibility of read lengths, OLC assemblers are also more robust to sequencing errors compared to de Bruijn graphs due to the fact that they look at maximum overlaps with high identity between reads and not just of size *k*-1, where a sequencing error could lead to a new branch in de Bruijn graph. After computing all pairwise overlaps, a multiple alignment is created and the consensus sequence is extracted. As mentioned earlier, repetitive sequences can lead to spurious overlaps; however CA has built in methods to deal with repeats and hopefully avoid false merges (Myers, Sutton et al. 2000).

OLC assemblers are not typically used with NGS data because they do not scale well with the number of reads obtained from these technologies. MaSuRCA has solved

this problem by assembling the paired end reads in to 'super-reads' and passing the super-reads to the Celera Assembler, along with the accompanying information about coverage for each super-read (Zimin, Marcais et al. 2013). To form a super-read, a read is extended on each end using a *k-mer* count lookup table and the extension stops when there is no longer a unique solution. Every read is contained in one super-read, though no super-read is fully contained in another. Additionally, super-reads can contain many reads, which provides drastic data reduction (Zimin, Marcais et al. 2013). The OLC algorithm is run on the super-reads, mate-pair reads and any additional long reads that can be supplied by the user.

To generate additional to reads to use as input to MaSuRCA, we used a string graph assembler (SGA) (Simpson and Durbin 2012) to pre-assemble subsets of paired-end reads from a fosmid library. The string graph is a slight variation on the OLC assembly algorithm (Figure 20C). Redundant information is removed by discarding reads that are entirely contained within another, which reduces the total number of vertices. Vertices are connected with bi-directional edges. An edge (*v,u*) leaving a vertex *v* and going into vertex *u* represents the sequence that precedes the start of the read sequence at vertex *u.* Similarly, the edge (*u,*v) represents the sequence of read *u* that comes after the end of read *v*. We chose the SGA assembler for the benefits of an overlap graph. This specific implementation uses compressed data structures (i.e. FM-index) so it can be run quickly and with minimal memory (Simpson and Durbin 2012). However, the computational time does not scale well as the number of reads increases which is why we only used it to pre-assemble subsets of reads and not for the entire assembly.

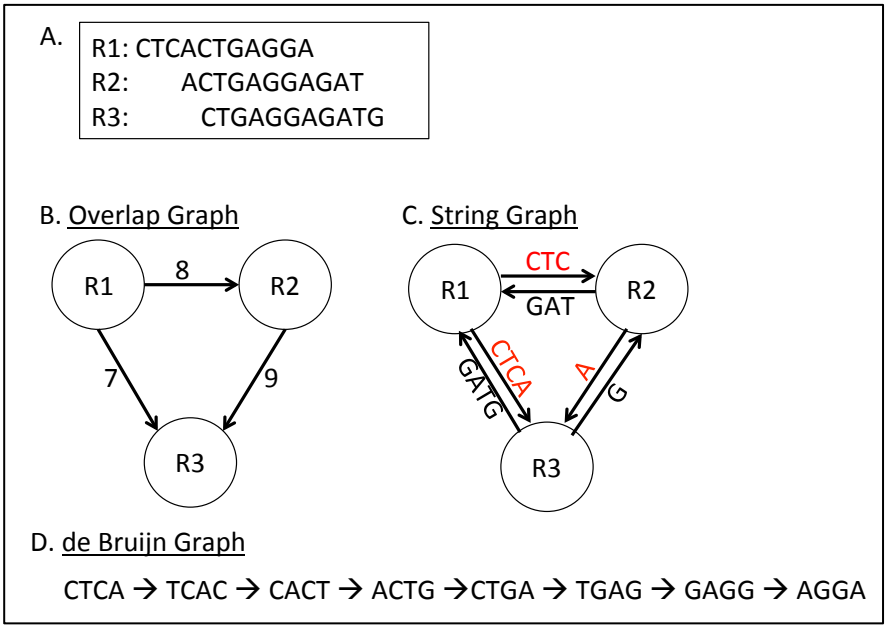


Figure 20. Graphical representations used in sequence assembly. The example reads that align as shown (A) could be represented by an overlap graph (B), string graph (C), or de Bruijn graph (D). The weights on the edges of the overlap graph correspond to the length of overlap (B). The edges leaving each node with red sequence above indicate the sequence from that read that comes before the node it is entering (C). The nodes pointing into a node with black sequence below represent the sequence that comes after the read at that node (C). The de Bruijn graph is an example where *k=4* (D).

ASSEMBLY STRATEGY

All sequencing of the humpback whale was done at the Genome Technology Center at the University of California Santa Cruz under the direction of Nader Pourmand. We prepared an 180bp paired end library and two mate-paired libraries with target insert sizes of 2Kb and 5Kb and sequenced each as 2x100bp reads (Table 7). Our initial assembly (*MegNov*.v01) was done with ALLPATHS-LG, but only used the 180bp fragment library and the 2kb mate-pair library as input. This was done primarily as a test, and our first real attempt (*MegNov*.v02) was also done with ALLPATHS-LG but made

use of the 5kb mate-pair library in addition to the other two libraries. The algorithm predicted that the genome size would be between 2.7 and 2.8 billion bases; however the resulting assembly was approximately 2.1Gb, suggesting we are missing a substantial portion of the genome. This result was partly due to the repetitive nature of the genome, which can collapse repeats, resulting in a smaller assembly. Another issue is that ALLAPTHS-LG highly recommends higher coverage for the paired end and mate-pair libraries in addition to at least one long insert library with an insert of greater than 10Kb. We had attempted to sequence a 10Kb mate-pair library and fosmid ends with a ~40Kb insert, but neither of these libraries were successful. However, we fragmented the fosmid library into ~425bp fragments and created a paired end library to add as input for the next assembly.

Library Type	**Fragment/Insert Length**	**# total reads**	**Coverage**
Paired-end	180bp	1.27 billion	45X
Paired-end	425bp (sheared fosmid clones)	500 million	100-200X*
Mate-pair	2Kb	326 million	10X
Mate-pair	5kb	269 million	9X

Table 7. Sequencing libraries used for the humpback whale genome. All libraries were used for the assembly of *MegNov.v03* using SGA and MaSuRCA. The *MegNov.v02* assembly did not include the paired-end 425bp fragment library that was created by sonicating fosmid clones. The *MegNov.v01* assembly only used the 180bp fragment library and the 2Kb mate-pair library. * The fosmid library coverage represents the estimated coverage for the fraction of the genome represented by the fosmids, not full genome coverage.

Our approach for improving the humpback whale assembly was three fold: (1) only use the highest quality input, (2) simplify the problem by pre-assembling smaller

pieces of the genome that we know belong together, (3) reduce the data to make our methods less computationally intensive. For the ALLPATHS-LG assembly, we had only removed the adapter sequences and poor quality reads were filtered by ALLPATHS. For this next assembly we invested a significant amount of time into pre-filtering our reads. This pre-filtering helps to eliminate errors in the assembly as well as decrease the number of input sequences to the assembler, which can speed up computational time. We removed all reads that did not pass the Illumina filter (indicated in the fastq header) and clipped poor quality bases from the ends of reads in addition to removing all adapter sequences and Illumina artifacts.

After all reads were filtered, the 2x100bp reads from the 180bp fragment library were joined where their ends overlap. Typically the quality of the sequence decreases toward the end of the read, so by joining reads, we are able to eliminate low quality regions (Figure 21). We were able to reduce the number of reads from 632,873,374 paired reads (i.e. over 1.2 billion individual reads) to 509,915,678 single end reads, with an average length of 157bp. Only reads that were successfully joined were used in the assembly.

To simplify the assembly problem, we pre-assembled contigs from fosmid sequence libraries. When the libraries were created, pools of approximately 50-100 fosmids were given a unique barcode so we were able to assemble each pool individually, knowing that the each pool assembly should be roughly 2-4Mb (i.e. the number of clones x 40Kb). We sequenced a total of 120 pools of fosmids and assembled each individually using the String Graph Assembler (Simpson and Durbin 2012) and kept all contigs that

were at least 200bp long which resulted in approximately 570Mb of assembled clones. Out of 1.06 million contigs, the average size was 535bp, with a maximum of 35,612bp. 10% of the contigs were at least 1Kb long and over 1,500 contigs were greater than 10Kb. The maximum read length that MaSuRCA can take as input is 2,047bp so all pre-assembled contigs were sheared to 2,047bp with an overlap of 1,500bp. This resulted in approximately 1.2 million long reads that now contain the information from over 500 million sequence reads from the fosmid library.

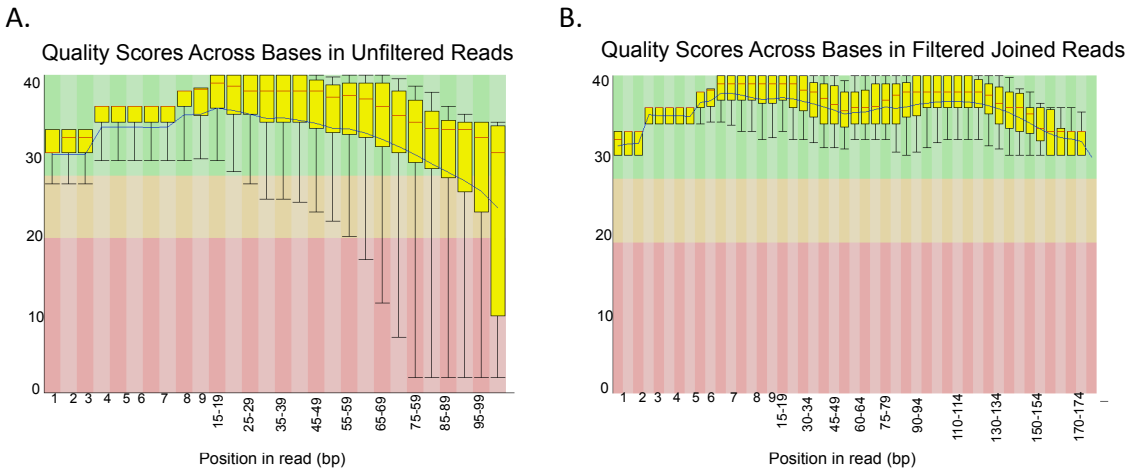


Figure 21. Base quality scores across reads. Before filtering and clipping of poor quality bases, the distribution of quality scores across the 100bp reads (from the 180bp fragment library) is highly variable and the quality drastically decreases toward the 3' end of the reads (A). After filtering poor quality reads, clipping adapters and low quality bases and joining the forward and reverse sequences where they overlap we see the base quality across the long reads increase and maintain consistency for the full length (B). The green indicates good quality base scores, red is poor quality scores and orange defines the intermediate qualities. These graphs were are produced by FastQC (Simon 2012).

MaSuRCA further reduces the data by assembling the paired end reads into super-reads before passing them to the adapted Celera Assembler based on OLC. We already reduced the number of paired end reads by joining the mates from the 180bp fragment library which condensed over 1.2 billion individual reads into 509 million reads. These reads, in addition to the unassembled fosmid reads, were assembled into 66.4 million super-reads. The super-reads have an average length of 305bp, a maximum of 29,585 and ~6% are over 1Kb. Super-reads greater than 2,047bp are sheared into 2,047bp fragments with an overlap of 1,500bp, as we also had done with the fomid contigs. The full assembly using MaSuRCA with the pre-assembled SGA fosmid clones was run on the Genepool cluster run by the National Energy Research Scientific Computing Cluster.

ASSEMBLY STATISTICS AND QUALITY ANALYSIS

After completing this new assembly, we sought to assess the contiguity, completeness and accuracy of our humpback whale draft genome. Most notably, we have improved the overall assembly size from the *MegNov.v02* assembly, which was approximately 2.1Gb to approximately 2.4Gb with *MegNov.v03* (Table 8). Both ALLPATHS-LG and MaSuRCA estimate the genome size based on *k-mer* frequencies; however, the algorithms are slightly different so the ALLPATHS assembler estimates the genome to between 2.7Gb and 2.8Gb and MaSuRCA estimates the genome to be between 2.2Gb and 2.3Gb. The exact reason for the discrepancy is unknown, but based on flow cytometry, the estimated genome size for some toothed whales (e.g. beluga whale, bottlenose dolphin and the Chinese River dolphin) suggest cetacean genomes are around 3 billion

base pairs (Gregory 2013). Therefore, when assessing the completeness using NG statistics we used an estimated genome size of 2.8Gb. We have summarized some of the assembly statistics in Table 8.

Some of the most frequently reported genome assembly statistics are the N50 and the NG50. The N*x* value is defined as the length at which *x*% of the assembled genome is contained in contigs (or scaffolds) greater than or equal to N*x*. NG statistics are very similar; however they are normalized to the estimated full genome size as opposed to the length of the assembly. Larger N-statistic (e.g. N50 and NG50) values reflect a higher degree of contiguity in the assembly. Our ALLPATHS assembly (*MegNov02*) has a scaffold N50 of 95.7Kb and a contig N50 of approximately 9.4Kb. If we use the genome estimate of 2.8Gb, the contig and scaffold NG50 values are 63Kb and 4.7Kb, respectively. It is obvious that the addition of the 5kb mate-pair library resulted in a large improvement in the scaffold length and overall contiguity compared to the *MegNov.v01* assembly (Table 8). The most recent version of the humpback whale genome (*MegNov03*) has not only improved on the total assembly length, but also shows significant improvements in contiguity which is reflected in the N50 and NG50 statistics (scaffold N50=144.2Kb; contig N50=11.2Kb; scaffold NG50=116Kb; contig NG50=8.8Kb) (Figure 22).

Though the *MegNov.v03* assembly contains many more scaffolds than the *MegNov02*, this is due to the fact that ALLPATHS-LG only outputs scaffolds (and contigs) that are at least 1Kb. The better comparison between the two is to only consider scaffolds that are at least this size for the MaSuRCA assembly. In doing this, we find that

98% of the assembly is contained in scaffolds greater than or equal to 1Kb, and therefore remains a significant improvement over the previous *MegNov02.v02* assembly.

	MegNov.v01	MegNov.v02	MegNov.v03
Software	ALLPATHS-LG	ALLPATHS-LG	MaSuRCA + SGA
Scaffold Length (bp)	1,829,725,285	2,093,296,260	2,388,269,474
Len. of Scaff. ≥ 1Kb (bp)	1,829,725,285	2,093,296,260	2,336,274,083
# Scaffolds ≥ 1kb	166,643	48,456	64,606
# total Scaffolds	166,643	48,456	182,196
Scaffold N50 (bp)	19,125	95,714	144,235
Scaffold NG50 (bp)	9,454	63,182	116,053
Scaffold L50 count	27,326	6,176	4,823
Scaffold LG50 count	63,225	10,731	6,413
Longest Scaffold (bp)	285,317	979,715	1,243,294
# Scaffolds > 1Kb	165,796	48,212	64,561
# Scaffolds > 10Kb	60,091	31,805	26,468
# Scaffolds > 100Kb	217	5,787	7,528
# Scaffolds > 1Mb	0	0	2
%Gaps	6.89%	11.56%	3.38%
Contig Length (bp)	1,703,724,128	1,851,304,040	2,307,471,974
Contig N50 (bp)	6,350	9,437	11,273
Contig NG50 (bp)	2,922	4,784	8,828
Contig L50 count	80,066	57,156	60,573
Contig LG50 count	206,067	127,279	85,239
Longest Contig (bp)	380,664	128,011	103,933
# total Contigs	380,664	320,806	529,825
# Contigs > 1Kb	375,824	317,618	301,933
# Contigs > 10Kb	32,705	52,238	72,505
# Contigs > 100Kb	0	3	2
%GC	40.81%	40.81%	40.74%
%AT	59.19%	59.19%	59.26
Ambiguities/10Kb	7.36	7.9	7.2

Table 8. Summary statistics for the humpback whale genome assemblies. NG50 and LG50 values are based on a genome size of 2.8Gb.

Additionally, the L statistics (L50 and LG50) are defined to be the number of contigs/scaffolds (greater than or equal to the N50) that contain 50% of the assembly (L50) or the estimated genome size (LG50). These values are lower in the *MegNov.v03* assembly, which means a larger portion of the genome is contained in a smaller number of scaffolds/contigs and therefore the assembly is on average more contiguous since this can only be explained by having some larger contigs and scaffolds than the previous assembly.

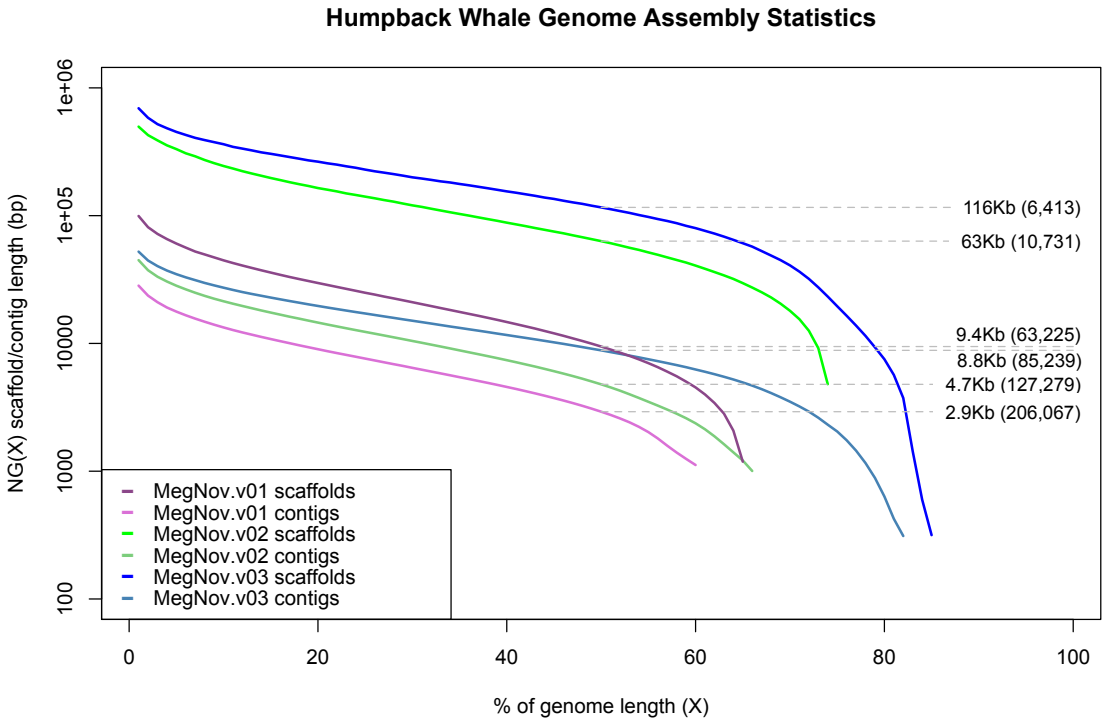


Figure 22. NG statistics of three humpback whale genome assemblies. The length of the genome is assumed to be 2.8Gbases. The x-axis values (*X*) are the % of the genome length and the y-axis represent the NG(*X*) statistics for the scaffolds and contigs of the three humpback whale assemblies. The NG50 (i.e. *X*=50) values are indicated on the right side of the graph and correspond to where the plotted data line intersects with the gray, hashed line. The numbers of scaffolds/contigs that account for 50% of the genome (i.e. the LG50) are shown in parentheses. The y-axis is plotted on a log scale.

Another major improvement in the most recent assembly is the drastic decrease in the percentage of the scaffolds that are gaps (i.e. N's). Almost 12% of the *MegNov.v02* humpback whale assembly is comprised of gaps and with the new *MegNov03* assembly, we have reduced the number of gaps to just over 3%. The majority of length that was gained in the *MegNov.v02* assembly relative to *MegNov.v01* was due to scaffolding with the additional 5kb mate-pair reads, so there are substantially more gaps. Gaps can inflate the scaffold N50 values, so it is important to also look at the total number of bases (A,C,G,T) to assess a genome. *MegNov.v03* is therefore a significant improvement over *MegNov.v02* because not only is it more contiguous as shown by the NG50, but it also contains nearly 500 million additional informative bases.

The three assemblies give similar values for the number of ambiguous bases as well as the GC content (Table 8). The number of ambiguous bases includes both sequencing errors as well as heterozygous single nucleotide polymorphisms (SNPs), which are inevitable in a diploid mammalian genome. Similar values are obtained from all three assemblies and are consistent with reported statistics from an individual human genome (6.15 heterozygous positions per 10Kb) (Levy et al. 2007). The average genome GC content, which is almost identical across all assemblies, is also consistent with other mammalian genomes (e.g. human ~41%) (Lander et al. 2001).

We assessed the overall repeat composition of the *MegNov.v03* assembly using the annotated mammalian repeats in RepBase with the software Censor (Kohany, Gentles et al. 2006). We find that 34.9% of the assembly is annotated as repetitive. This is less than we would expect and is likely due to collapsed repeat regions. There are only 6,695

instances of the 1.7Kb common cetacean element, which only makes up 0.06% of the assembly. Based on previous estimates that this one repeat accounts for 15% of the genome, there should be approximately 250,000 copies the full genome totaling ~420Mb. This could explain the missing ~400Gb of our assembly compared to the estimated size of 2.8Gb. 8,118 instances of the heavy satellite are annotated by Censor in *MegNov.v03,* which also accounts for less than 1% of the genome (0.07%). We observe that these annotations are frequently found on short scaffolds (~500bp), which demonstrates the fragmentation that occurs with tandem repeats. L1 repeats make up 44.6% of all annotated repeats in the humpback whale genome and account for 18.75% of the total assembly (Figure 23). The second most common class of repeats are the SINEs which total approximately 9% of the genome if we combine those annotated as 'SINE" and "SINE2/tRNA" and ~20% of the total repeats.

The *MegNov03* assembly is comparable to the published bottlenose dolphin (*Tursiops truncatus*) genome (*Ttru14*), which was done with low-coverage (2.6X) Sanger sequencing at Baylor College of Medicine and assembled by the Broad Institute. The dolphin in 240,900 scaffolds span 2.5Gb with a scaffold and contig N50 of 109Kb and 11.8Kb, respectively (as reported by Ensembl, release 73). As additional cetacean genomes are published we will be able to do more robust comparative genomic studies in addition to working to improve our assembly to achieve higher contiguity and find ways to handle repeat regions more effectively.

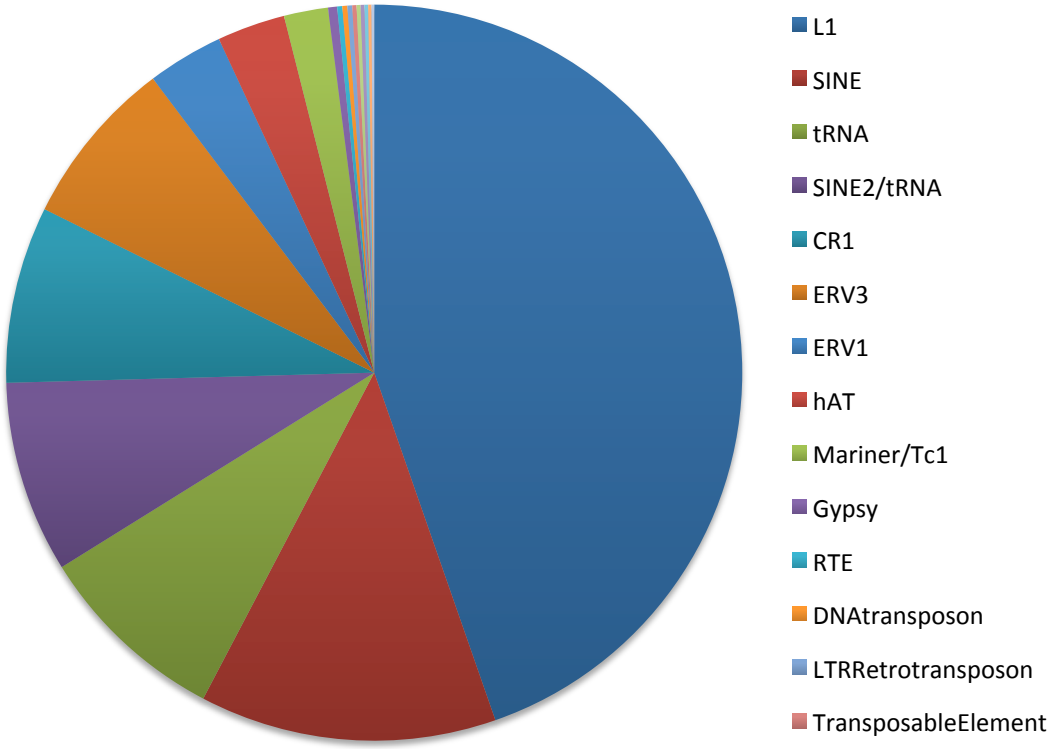


Figure 23. Distribution of Repeat Elements in the Humpback Whale Genome. Repeats were annotated with Censor in the humpback whale assembly *MegNov.v03* based on the mammalian library provided by RepBase (Kohany, Gentles et al. 2006). The majority of the repeats are L1 LINEs (large blue wedge) accounting for 44.6% of all the repeats. The most abundant repeats are listed in the legend and go clockwise starting from the large L1 wedge.

Obviously improving the length and contiguity of a genome assembly is not enough; it is also imperative that the assembly is accurate. To get a general idea of they accuracy, we directly compared our assemblies (*MegNov.v02* and *MegNov.v03*) to regions of the humpback whale genome that have been independently sequenced by Sanger sequencing and deposited in GenBank. There are currently 2,351 nucleotide

sequences in GenBank that exists for this species. We ran BLAST to query them against the *MegNov.v02* and *MegNov.v03* assemblies to get a measure of accuracy. We find that 97.7% of the sequences are found in the *MegNov.v02* assembly and 96.6% (only 27 fewer genes) produce a hit in *MegNov.v03*. Though there are slightly fewer hits in the *MegNov.v03* genome, the hits have higher percent identities on average (98.43% vs. 95.44%). The distribution of all top hit percent identity scores for the two assemblies are shown in Figure 24, in which we clearly see hundreds of sequences that are aligning at a much lower identity in the *MegNov.v02* assembly.

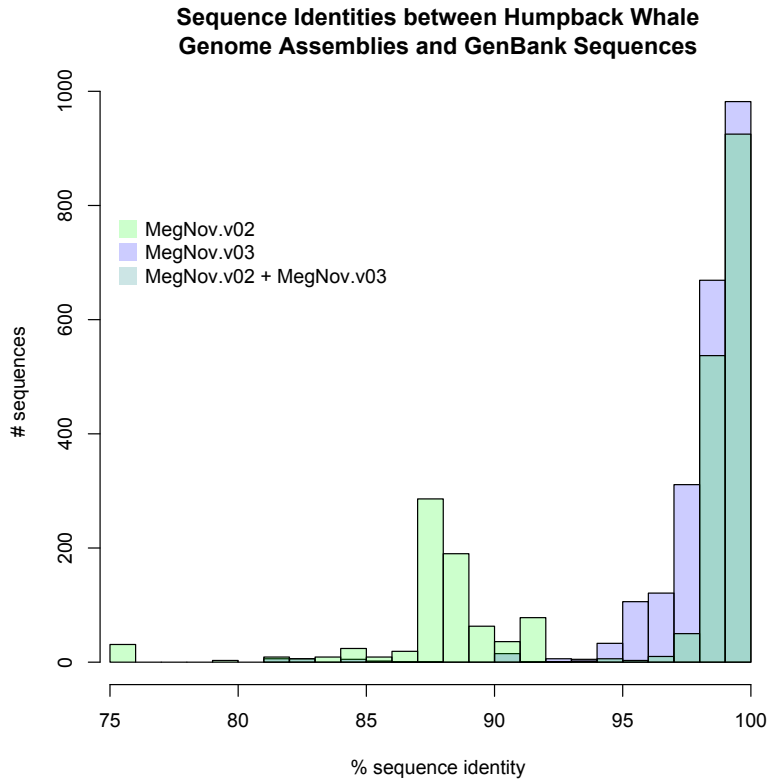


Figure 24. Distribution of sequence identities from BLAST alignments. BLAST was used to align the humpback whale draft assemblies to previously (Sanger) sequenced humpback whale sequences in GenBank. The sequences from GenBank generally align better with the *MegNov.v03* assembly compared to the *MegNov.v02*. The blend of the two colors defined in the legend represents the overlapping regions of the histograms.

As an additional test for assembly accuracy, we used gene transcripts that were independently assembled to determine their agreement with the genomic sequence, though in the case of disagreement it is not clear which sequence is likely to be correct. We sequenced the RNA that was isolated from the same skin biopsy of the whale and *de novo* assembled the transcripts with the Trinity pipeline offered by the Broad Institute (Grabherr et al. 2011). GenBlastG (She et al. 2011) was used to annotate the top 500 longest predicted open reading frames (ORFs) in the *MegNov.v03* assembly. We are able to successfully annotate at least part of all 500 genes; however short annotations may be a result of conserved protein domains. A more accurate representation is the number of genes that are annotated with good coverage (≥ 70% of the query ORF). We find that 86.8% have at least 70% coverage on one scaffold and over half of these genes and have an average percent identity greater than 91%. Over half of the genes with quality coverage align with the predicted ORFs with greater than 98% identity. These results provide confidence in our assembly in addition to validating the predicted ORFs from the assembled transcript data.

Though our overall goal was the assembly a complete humpback whale genome, many downstream analyses will focus heavily on the gene sequences. We used the software CEGMA (Parra et al. 2009) to estimate the completeness of our assembled genic regions. CEGMA uses a set of 248 genes that are conserved across eukaryotes, but with varying degrees of sequence identity. The core eukaryotic genes (CEGs) are divided into four categories based on their sequence conservation across eukaryotes: low, medium-low, medium-high and high. A well-assembled genome would contain a high fraction of genes in each category, whereas an incomplete genome would contain a low percentage

of genes in the four groups (Figure 25). In a well-assembled divergent genome (e.g. protozoans *T. gondii* and *P. falciparum*), genes that are most highly conserved would still be found while those that are less conserved would be harder to identify.

We find that our humpback whale genome assemblies (*MegNov.v02* and *MevNov.*v03) both cluster with the complete genomes such as chimp (Figure 25); however, the *MegNo.v03* assembly does slightly better overall (Table 9). The total number of CEGs found in the two assemblies is almost identical, but 12 additional complete CEGs are found in the *MegNov.v03* assembly, which is a 5% increase in complete genes. The number of CEGs that can be annotated is highly correlated with the total number of genes in the genome (Parra, Bradnam et al. 2009) so we estimate that our *MegNov.v03* assembly should contain approximately 87% of the full set of humpback whale genes.

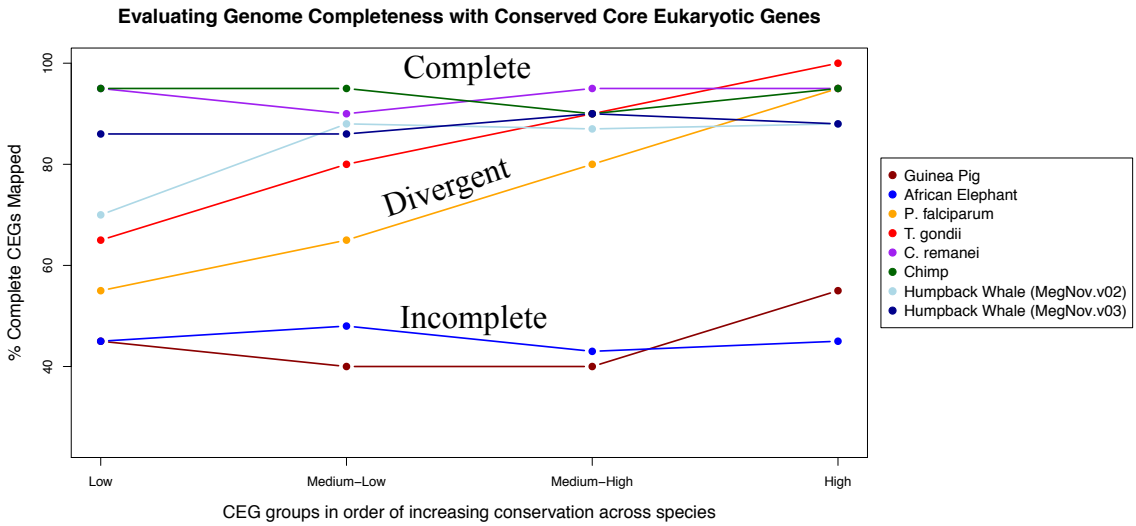


Figure 25. Evaluation of genome completeness with conserved core eukaryotic genes. The *MegNov.v03* assembly is slightly more complete than the previous whale assembly based on the annotation of CEGs across variable levels of conservation. This figure was adapted from (Parra, Bradnam et al. 2009) to include our humpback whale data.

Assembly	# Complete	% Complete	Total CEGs	% Total CEGs
MegNov.v02	205	82.66%	241	97.18%
MegNov.v03	217	87.50%	242	97.58%

Table 9. Core eukaryotic genes found in the humpback whale genome assemblies. CEGMA defines a gene to be complete within a genome if it is at least 70% of the query gene length. All other found core eukaryotic genes (CEGs) annotated in a genome are defined to be partial hits and the total number of CEGs (out of 248) that are found in a genome include the complete and partial annotations.

MATERIALS AND METHODS

Tissue Collection

The tissue used for genomic sequencing was collected from a female humpback whale (Salt) off the coast of Cape Cod using biopsy techniques (Lambertsen 1987, Palsboll et

al. 1991) and immediately frozen in liquid nitrogen. The collection was organized by Dr. Per Palsboll, at the University of Groningen, and Dr. Jooke Robbins, at the Provincetown Center for Costal Studies.

DNA Extraction

Female humpback whale skin tissue was used for high molecular weight genomic DNA isolation following the standard protocol for the DNeasy Blood & Tissue purification kit (Qiagen).

DNA Library Construction and Sequencing

All sequencing libraries were prepared at the Genome Technology Center at the University of California Santa Cruz under the direction of Nader Pourmand. The paired-end library was prepared with ~1µg of genomic DNA that had been sheared using the Covaris S2 sonicator. Quality and quantity of fragmented genomic DNA was confirmed using an Agilent Bioanalyzer 2100 DNA High Sensitivity chip. A standard ~180bp insert size paired-end library was constructed using TruSeq DNA sample preparation kit (Illumina). Final size selection between 290-310bp was done using Caliper LabChip XT DNA 750 chip following manufacturer's recommendation. The quality and functionality of the library was confirmed by sequencing in a single HiSeq 2000 lane. After confirming the quality, library was subjected to 2X100bp paired-end sequencing in three lanes of HiSeq 2000.

Two mate pair libraries (2.0Kb and 5.0kb) were prepared using the SOLiD Mate Pair Library Preparation Kit with modifications. Approximately, 10μg genomic DNA was fragmented using the S2 Adaptive Acoustic Device (Covaris). For the 2Kb mate-paired library, size selection of DNA fragments (1.8-2.5Kb) was carried out through LabChip XT, developed by Caliper Life Sciences. For the 5Kb size selections, sheared DNA was electrophoresed in a 0.7% agarose gel for ~16 hours. A band corresponding to ~4.5-5.5kb was excised from the gel and purified by ethanol precipitation. The size selected fragmented DNA was subjected to end repair followed by ligation of 5' dephoshorylated biotinylated MP adaptor on both sides of the DNA fragments. The MP adaptor ligated DNA was circularized through intra-molecular hybridization at very low concentration, which results in a nick at the 3' ends of the internal adaptors. In order to generate ~100bp paired tags, nicks in the circularized DNA were bidirectionally extended into the insert DNA using timed nick translation reaction by DNA polymerase I. The nick translated DNA was digested with T7 Exonuclease and S1 Nuclease to release the tags. These tags were subjected to end repaired, adenylation and then ligated with Illumina non-multiplexed adaptors. The final DNA molecules were bound to streptavidin beads. The library DNA molecules bound with beads were enriched by standard Illumina non-multiplexed paired-end primers. Final amplified libraries were analyzed using an Agilent Bioanalyzer 2100 DNA High Sensitivity chip and size selection was done using a Caliper LabChip XT DNA 750 chip following manufacturer's recommendation. Both mate-paired libraries were subjected to paired-end sequencing in two lanes of HiSeq 2000 sequencer generating 2×100bp reads.

The whale fosmid library was constructed by using the pCC2FOS Fosmid vector (CopyControl™ Fosmid Production Kit; Epicentre Technologies). Approximately, 20µg high molecular weight genomic DNA was end-repaired and purified by ethanol precipitation. The end repaired DNA was ligated to ready to clone pre-linear pCC2FOS fosmid vector according to manufactur's specifications (10:1 molar ratio of vector and insert DNA). Fosmid clones were packaged using MaxPlax Lambda Packaging Extract and stored at 4°C in 1 ml of Phage Dilution Buffer according to manufacturer's instructions. Primary packaged Fosmid libraries were transfected into EPI300-T1R Phage T1-resistant E. coli Plating strain and incubated for 1 hour at 37°C. The infected bacteria were spread at high density on LB plates containing 12.5µg/mL chloramphenicol and incubated at 37°C for 18 h. Random bacterial colonies were selected for bulk infection in 10ml LB with 10mM MgSO4 and 0.2% maltose in a shaking incubator for 14 hours at 37°C. Fosmid DNA was extracted using the Qiagen MiniPrep Kit following the manufacturer's protocol and individually sheared using Covaris S2. Size selection of DNA fragments (400-450bp) was carried out through an automated electrophoretic DNA fractionation system, LabChip XT, developed by Caliper Life Sciences. Multiplexed paired-end sequencing libraries were prepared using TruSeq DNA Sample Preparation Kit. 120 fosmid pools (with roughly 50-100 fosmid clones per pool) were subjected to paired-end sequencing in HiSeq 2000 one lane generating 2×100bp reads.

RNA isolation and purification

Total RNA was extracted with QIAzol Lysis Reagent, purified on RNeasy spin columns (Qiagen), and the RNA integrity and quantity was determined on the Agilent 2100 Bioanalyzer (Agilent) with the manufacture's protocol. The total RNA was treated with DNase using DNase mix from RecoverAll™ Total Nucleic Acid Isolation kit (Applied Biosystems/Ambion) and subjected to cDNA synthesis.

RNA-seq Library

The RNA library was prepared and sequenced by the Genome Technology Center at the University of California Santa Cruz under the direction of Nader Pourmand. The Ovation RNA-Seq system V2 (Nugen) was used for cDNA synthesis and RNA amplification was performed as described in detail in published literature (Tariq, Kim et al. 2011). Briefly, the total RNA was reverse transcribed to synthesize the first-strand cDNA by using a combination of random hexamers and poly-T chimeric primer. Double-stranded DNA was generated by fragmentation of the mRNA template strand using RNA-dependant DNA polymerase. The dsDNA was purified using Agencourt RNAClean XP beads. The DNA was amplified linearly using a SPIA process in which RNase H degrades RNA in DNA/ RNA heteroduplex at the 5'-end of the double-stranded cDNA, after which the SPIA primer binds to the cDNA and the polymerase starts replication at the 3'-end of the primer by displacement of the existing forward strand. Finally, random hexamers were used to amplify the second-strand cDNA linearly, as described previously (Tariq, Kim et al. 2011).

The double-stranded cDNA obtained after the Ovation V2 RNA-Seq system, 0.5–1 μg of double-stranded DNA was used for library. The cDNAs were sheared down to 350-450bp using the Covaris S2. A target insert size of 350-450bp was then size-selected using an automated electrophoretic DNA fractionation system, LabChip XT (Caliper Life Sciences) and paired-end sequencing libraries were prepared using Illumina's TruSeq DNA Sample Preparation Kit. Following library construction, samples were quantified using the Agilent Bioanalyzer per manufacturer's protocol. Libraries were sequenced using the Illumina HiSeq 2000 with sequencing paired-end read length at 2 x 100bp. Reads were de-multiplexed using CASAVA (version 1.8.2).

Read Filtering

For the ALLPATHS-LG assembly, Illumina and Solid adapter sequences were clipped from the ends of reads using fastq-clipper from the ea-utils package (Aronesty 2011). For the MaSuRCA and SGA assembly, all reads were filtered using fastq-mcf in the ea-utils 1.1.2-484 package (Aronesty 2011). The adapter sequences were clipped from the ends and known Illumina artifacts were removed (a FASTA file of artifacts was provided by the Department of Energy Joint Genome Institute). The cloning vector sequence and *E. coli* contamination was also removed from the fosmid library reads. Poor quality bases were clipped from the ends of reads (score cutoff=10) and reads less than 30 bases were discarded after these filtering steps. Reads were maintained as mates, so both were discarded if one did not pass the filters. Filtered reads from the 180bp fragment library were joined using fastq-join in the ea-utils package (Aronesty 2011). Only pairs that were successfully joined were used as input into the MaSuRCA assembly.

Assembly of MegNov.v01

The very first assembly (*MegNov.v01*) was run with ALLPATHS-LG provided by the Broad Institute (Gnerre, Maccallum et al. 2011). We used the default settings and set the ploidy equal to 2, the expected genome size to be 3Gb and the minimum contig size for reporting to be 1,000bp. The 2x100bp reads from the 180bp fragment library and the 2kb mate-pair library were used as input after clipping Illumina and Solid adapter sequences. The assembly was initialized by John St. John at the University of California Santa Cruz on a server with 1TB of RAM and 64 cores. It required over 524GB of RAM and took approximately one month to finish.

Assembly of MegNov.v02

The *MegNov.v02* assembly was run using ALLPATHS-LG assembler provided by Broad Institute (Gnerre, Maccallum et al. 2011). We used the default settings and set the ploidy equal to 2, the expected genome size to be 3Gb and the minimum contig size for reporting to be 1,000bp. The 2x100bp reads from the 180bp fragment library and the 2kb and 5kb mate-pair libraries were used as input after removing any remaining adapter sequences. John St. John, at the University of California Santa Cruz, ran this assembly on a server with 1TB of RAM and 64 cores. It required 630GB of RAM and took approximately one month to complete.

Assembly of MegNov.v03

The most recent draft of the humpback whale genome (*MegNov.v03*) used both the SGA and MaSuRCA assemblers (Simpson and Durbin 2012, Zimin, Marcais et al. 2013). Fosmid contigs were pre-assembled with SGA with the parameters specified in the example bash script in the Appendix. Contigs that were greater than 2047bp long were sheared into 2,047bp long fragments with an overlap of 1,500bp using the program *splitter,* written by Gary Williams as part of EMBOSS (Rice et al. 2000). These were assigned quality values using the script provided by PBJelly (English et al. 2012) and converted into *frg* files for input into MaSuRCA. The 180bp fragment library joined reads along with the filtered 2kb and 5kb mate-pair libraries were also used for this assembly. MaSuRCA was run on the Genepool cluster at the Department of Energy. Its peak memory usage was less than 300GB, and though the contigs were assembled in ten days, the scaffolding took approximately 40 additional days.

De Novo Transcript Assembly from RNA-seq Reads

The Trinity pipeline provided by the Broad Institute (Grabherr, Haas et al. 2011) was used to assemble the raw RNA-seq reads into putative transcripts. From approximately 142 million paired reads (i.e. ~284 million individual reads), we assembled 156,162 transcripts. We then used a Perl script that is part of Trinity (transcripts_to_best_scoring_ORFs.pl) to predict open reading frames (ORFs) in the transcripts. The longest ORF from each transcript was saved and translated into the corresponding amino acid sequence. This gave approximately 10,000 high-confidence

protein predictions. We used BLAST to align these protein sequences to the non-redundant Swissprot database. Requiring an E-value of less than or equal to 0.001, 1,824 transcripts remained. Predicted whale amino acid sequences were required to be +/- 15% of the length of the subject protein in the database. This eliminates significant hits due to highly conserved domains occurring in genes that are not true orthologs. 1,233 transcripts passed this final filter and were used for downstream analyses.

Genome Statistics and Analysis

Statistics found in Table 8 were generated with the script provided by Assemblathon2, which we adapted to properly split scaffolds into contigs (Bradnam et al. 2013). We ran the script on the scaffold files of each assembly and split on gaps (i.e. N's). The values we obtained were in agreement with those that overlap the statistics provided in the output by the assembler software. Repeats were annotated with Censor using the mammalian repeat library provided by RepBase (Kohany, Gentles et al. 2006).

The assemblies were compared to the humpback whale nucleotide sequences available in GenBank, which were all done with Sanger sequencing. We ran *blastn* using the 2,351 GenBank sequences to query the *MegNov.v02* and *MegNov.v03* assemblies. The top hit of each query was used to compare the average percent identities and the distributions were plotted in R.

We also compared the assembly to the RNAseq transcripts that we assembled with Tritinty (Grabherr, Haas et al. 2011). The top 500 longest open reading frame

predictions were translated to amino acid sequences. These were used as the query inputs to genBlastG to annotate these genes (based on homology) in the *MegNov.v03* assembly (She, Chu et al. 2011). We analyzed the number of genes with a hit, as well as the coverage and percent identity.

The open source software CEGMA (Parra, Bradnam et al. 2009) was used in order to access the completeness of the genome. We executed this using *mam* parameter was used which is optimized to work on mammalian genomes.

DISCUSSION

We have successfully assembled the genome of the female humpback whale known as Salt. After three assembly attempts we have continued to improve the assembly length, contiguity, completeness and accuracy. Data quality and data reduction were the key to the success of the *MegNov.v03* assembly. Without data reduction we would not have been able to take advantage of long pre-assembled reads to help overcome the highly repetitive nature of the genome. Unfortunately we were not able to successfully perform end sequencing, which would provide mated reads 30-40Kb apart and span many repeat regions. If we had been able to do this and also provide more coverage of the other libraries, ALLPATHS-LG would likely have had improved results compared to the *MegNov.v02* assembly. However, given our data, the flexibility of MaSuRCA allowed us to use long joined reads as well as pre-assembled fosmid contigs with an overlap-consensus-layout assembler.

Though there tends to be a lot of emphasis on statistics including N50, and total scaffold length, there are many other features of an assembly that determine its quality. The number of informative bases (i.e. the scaffold length minus gap length) is crucial, and it is also necessary that there is confidence in the order in which these bases are assembled. We have found that our *MegNov.v03* genome assembly contains nearly 97% of the humpback whale Sanger sequences deposited in GenBank and the sequences agree with high percent identity. This is extremely promising because many of these sequences in the database are targeting variable regions of the genome in order to distinguish individual whales. Our CEGMA analysis is consistent with these results and also predicts that the genome contains 97% of the genic regions (both partially and completely assembled) and is estimated to have 87.5% of the genes assembled in a complete form. This in in close agreement with what we find by annotating the predicted ORFs from the *de novo* assembled RNAseq transcripts so we are confident that this is a good estimate of the fraction of well-assembled genic regions.

Our analysis demonstrates with a high level of confidence that the humpback whale genome has high enough quality to use in down stream analyses. We are now able to use this genome to address questions surrounding Peto's paradox, which will be discussed in detail in Chapter 7. Multiple research groups have been pursuing genome-sequencing projects of other cetaceans. Our humpback whale genome is one of the first; however, the minke whale genome publication was just released and the bowhead whale should follow shortly. Additionally, there are groups focusing on toothed whales including the harbor porpoise, killer whale, narwhal and the sperm whale. The release of this genome enables comparative genomic studies to be conducted and allows our group

to focus on determining how evolution has shaped cancer suppression in humpback whales with comparison to the African elephant genome, in order to inform future cancer prevention strategies.

CHAPTER 7: GENOMIC ANALYSIS OF CANCER SUPPRESSION

INTRODUCTION

The genomic sequence of a species is the result of millions of years of evolution acting through both drift and natural selection. By analyzing the genome of an organism, and comparing it to others, we can gain insight into the process of evolution that gave rise to the current state of that organism. Given the theory that cancer risk should increase with body size and lifespan, natural selection has played a critical role in shaping the mechanisms of cancer suppression in large, long-lived animals to allow them to overcome this burden. The signature of this selection should be evident within their genomes. As we have shown with the African elephant, one possible way to combat an increased cancer risk is by duplicating tumor suppressor genes (e.g. *TP53*). In this study we investigate our genome assembly of the humpback whale (*MegNov.v03*) in search of duplicated tumor suppressor genes in addition to evidence of convergent and accelerated evolution of known tumor suppressors.

Convergent evolution describes the observation that similar features can arise independently in distant clades. On a molecular level, convergent proteins evolve from divergent amino acid sequences into more similar sequences along independent branches of life (Zhang and Kumar 1997). If tumor suppressor genes were evolving to have an enhanced or specialized function in large, long-lived organisms, we would expect the amino acid sequences of distantly related animals with similar body size and longevity to

look more similar to each other than we would expect based on our knowledge of species evolution (Figure 26). An analysis of the *Prestin* gene, which is thought to increase sensitivity and selectivity to high frequency sounds, reveals that dolphins cluster closely with echolocating bats in a phylogeny based on this amino acid sequence (Liu et al. 2010). This is an example of convergent evolution at the molecular level. The results suggest that the evolution of this gene in dolphins and echolocating bats resulted in similar protein sequences, and therefore possibly more similar functions, despite having different initial sequences in their more recent ancestors. We have surveyed tumor suppressor genes across three mammalian clades (show in Figure 26) in search of gene trees that show the elephant and whale clustering together instead of with the species they are more closely related to evolutionarily (Figure 26).

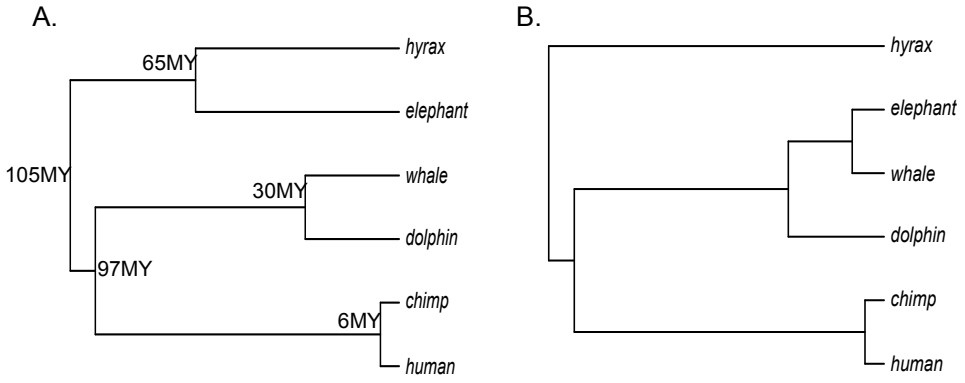


Figure 26. Illustration of convergent evolution. The species phylogeny of the two large, long-lived species under investigation (elephant and whale) along with their closest relatives that have been sequenced (hyrax and dolphin) and human and chimp is depicted on the left (A). Time since divergence is marked on the internal nodes. If there has been convergent evolution for large, long-lived bodies, then the amino acid sequences of the genes might produce a gene tree where elephant and whale cluster together (B).

Additionally, accelerated evolution of specific genes or gene sets associated with a single lineage is often a sign of a lineage-specific adaptation. The rate of protein evolution is conventionally measured using the ratio *dN*/*dS* (referred to as ω), where *dN* is the number of non-synonymous changes per non-synonymous site and *dS* is the number of synonymous changes per synonymous site (Miyata and Yasunaga 1980). If non-synonymous and synonymous substitutions evolved at a similar rate, ω would be equal to 1, and this is considered neutral evolution. If this ratio (ω) is significantly greater than one it implies that the protein is under positive selection whereas a ratio (ω) significantly less than one is considered purifying selection (Miyata and Yasunaga 1980). The majority of protein coding genes are under strong selective pressures, where synonymous substitutions occur at a higher frequency per site than non-synonymous, consequently, most protein coding genes in mammalian genomes are under purifying selection (Ohta 1973).

To test for accelerated evolution, we are interested in comparing the estimated ω's along different lineages to determine if a specific branch, or set of branches, have a higher ω value than the rest of the tree and therefore may be evolving at an increased rate. Thus, for accelerated evolution, ratio ω in a specific lineage does not have to be greater than one, just greater than the other branches. Previous work has shown that conserved non-coding elements near neuronal genes (Prabhakar et al. 2006) as well as genes involved in nervous system development have evolved at accelerated rates in the human genome (Dorus et al. 2004), which suggests these genes are in part responsible for the large, complex brain which has evolved in *Homo sapiens*. In this study we investigate three possible evolutionary scenarios (redundant genes, convergent evolution and

accelerated evolution) to elucidate the tumor suppressive mechanism(s) that have evolved in the humpback whale resulting in a decreased cancer risk relative to what we would predict based on their size and lifespan.

COPY NUMBER OF TUMOR SUPPRESSOR GENES

We previously analyzed the copy number of tumor suppressor genes in publically available mammalian genomes (discussed in Chapter 4), so we applied the same techniques to the humpback whale genome in search of redundant TSGs that may add robustness to their tumor suppressive pathways. We did not find any genes with extreme amplification in the humpback whale genome; however there are four tumor suppressor genes that appear to have one extra copy (i.e. two copies total) (Table 10).

Gene	Total Copy #	Description
TUSC2	2	1 retrogene, 1 full gene copy
CDK2AP2	2	1 expressed retrogene, 1 full gene copy
BECN1	2	2 full gene copies
TCEAL7	2	2 full gene copies

Table 10. Copy number of tumor suppressor genes in the humpback whale genome. *TUSC2* and *CDK2AP2* have retrotransposed genes as their additional copy and these are conserved in human. *BECN1* and *TCEAL7* have additional full copies of the genes, including introns and exons.

Two of the gene duplications that we find have annotations consisting of a single exon, unlike the top ranked gene model, and are therefore likely a result of

retrotransposition. *TUSC2* (tumor suppressor candidate 2), also known as *FUS1*, if often mutated or deleted in human cancers (Ji and Roth 2008) and normally functions to promote p53-dependent apoptosis (Deng et al. 2007). The human genome has a known processed pseudogene of *TUSC2* on the Y-chromosome; however the expression and function are unknown.

We also believe that the single exon additional gene copy we annotated for *CDK2AP2* is a known retrogene. *CDK2AP2* (cyclin-dependent kinase 2 associated protein 2) has three processed pseudogenes in the human genome (*CDK2AP2P1*, *CDK2AP2P2* and *CDK2AP2P3*), which are all expressed in multiple tissues including brain, heart, liver, breast and colon, with the highest expression in testes, but has no reported function (Bu et al. 2012). We do not predict that the processed copies of *TUSC2* and *CDK2AP2* are likely involved in solutions to Peto's paradox because we also observe processed copies in smaller mammals, such as humans.

BECN1 also has two retrotransposed gene copies in the human genome. Interestingly, our annotation of the humpback whale finds two intron-containing copies of *BECN1*, and no processed copies that pass our filters. These genes are located on two different scaffolds; however one of the scaffolds is 7,519bp and just coding sequence of *BECN1* spans 7,062bp so we cannot place the gene into context because it consumes the entire scaffold. The other scaffold containing a full gene model of *BECN1* is 58,376bp long and *BECN1* is at the very 5' end. The two sequences are conserved between copies and because the assembler did not collapse these into one sequence, it is likely that the coverage depth was informative of a duplicated region, though we cannot rule out the

possibility of an assembly error. The protein product of this gene promotes autophagy (i.e. the degradation of cellular components) and is associated with inhibition of cellular proliferation (Liang et al. 1999). This gene is mutated in multiple cancer types (most commonly in breast and ovarian cancer) and is haploinsufficient (Qu et al. 2003), so a redundant copy would protect the gene's function and require that more than one deleterious mutation occur to result in a phenotypic change.

Additionally we found a duplicate copy of *TCEAL7*, which only has one coding exon, but both copies include non-coding exons as well as introns. Like the *BECN1* gene, one copy of *TCEAL7* is found on a short scaffold that contains only the gene and is highly conserved with the top ranking gene model that is found on a larger scaffold. The genic region contains repetitive LINE elements, which may have contributed to the fragmentation of the assembly and placement of the duplicate copy on a short (1,257bp) unplaced scaffold. Nonetheless, *TCEAL7* is a negative regulator of the NF-kappa-B signaling pathway, which is often overexpressed in malignant cells (Arlt and Schafer 2002, Karin 2006, Rattan et al. 2010), so an additional copy may help to regulate this more consistently. NF-kappa-B is a transcription factor involved in a wide variety of functions including inflammation, cell growth, differentiation and apoptosis (Gilmore 2006, Karin 2006).

Due to the fact that the humpback whale genome assembly is comprised of unplaced scaffolds, we cannot put the suspected duplicate copies in genomic context because each copy is found on a different scaffold. Additionally, we queried the predicted

ORFs of our *de novo* assembled RNA transcripts from the humpback whale biopsy, but did not find evidence for expression of any of these four genes in the skin.

CONVERGENT EVOLUTION OF TUMOR SUPPRESSOR GENES

Next we sought to determine if any tumor suppressor genes show signs of convergent evolution along the elephant and whale lineages. We used amino acid sequence alignments to construct gene trees for 501 tumor suppressor genes using six species: human, chimp, dolphin, humpback whale, African elephant and hyrax. The elephant and whale were chosen as our species of interest with regard to Peto's paradox and we wanted to compare these to the known human tumor suppressor genes. The other three species were chosen as the most closely related sequenced species to the elephant, human and whale so we could determine changes that were specific to the large, long-lived animals (i.e. African elephant and humpback whale). Gene trees that appeared to be a result of a bad alignment after manual inspection were discarded.

After filtering, one gene that gave the desired phylogeny (Figure 26) remained, *UBE2D1,* which codes for a ubiquitin-conjugating enzyme. The maximum likelihood phylogeny shows elephant and whale clustering closely together and the other four species as a separate clade (Figure 27A). We find that the amino acid sequence of *UBE2D1* is perfectly conserved across 30 other mammals with sequences annotated in GenBank; however the first eight amino acids in the predicted elephant and whale sequences differ from all of the other mammals, yet are similar to each other (Figure 27B). These amino acid changes lie within the N-terminal alpha-helix 1, which is part of

the catalytic core domain (along with the L1 and L2 loops) and is responsible for E3 recognition and binding (Figure 28A) (Ye and Rape 2009, Kar et al. 2012). Due to the extreme conservation observed across other mammals, this finding is arguably an example of parallel evolution because both elephant and whale likely evolved from the same, or similar, ancestral amino acid sequences into their current sequences. When the starting points are the same, this is often referred to as parallel evolution as opposed to convergent evolution (Zhang and Kumar 1997).

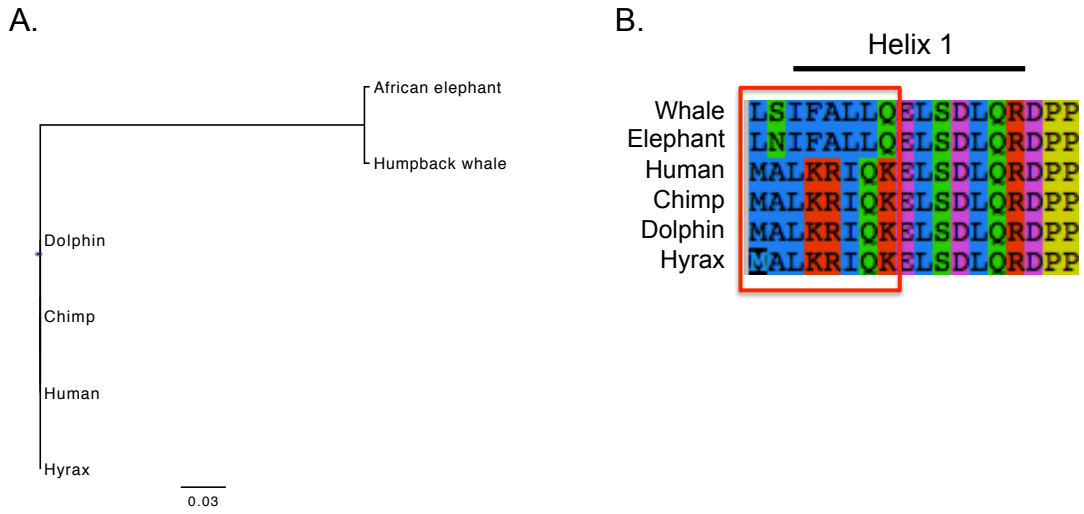


Figure 27. Evidence of convergent evolution of the UBE2D1 protein. The maximum likelihood gene of the UBE2D1 amino acid sequences shows the African elephant clustering with the humpback whale, though likely derived from the same starting sequence seen in other clades, which suggests parallel evolution (A). The first 8 residues in the elephant and whale sequences of UBE2D1 (outlined by a red rectangle) are more similar to each other than the other species, as shown in the multiple alignment (B). The position of alpha-helix 1 is indicated above the alignment and contains most of the altered amino acids.

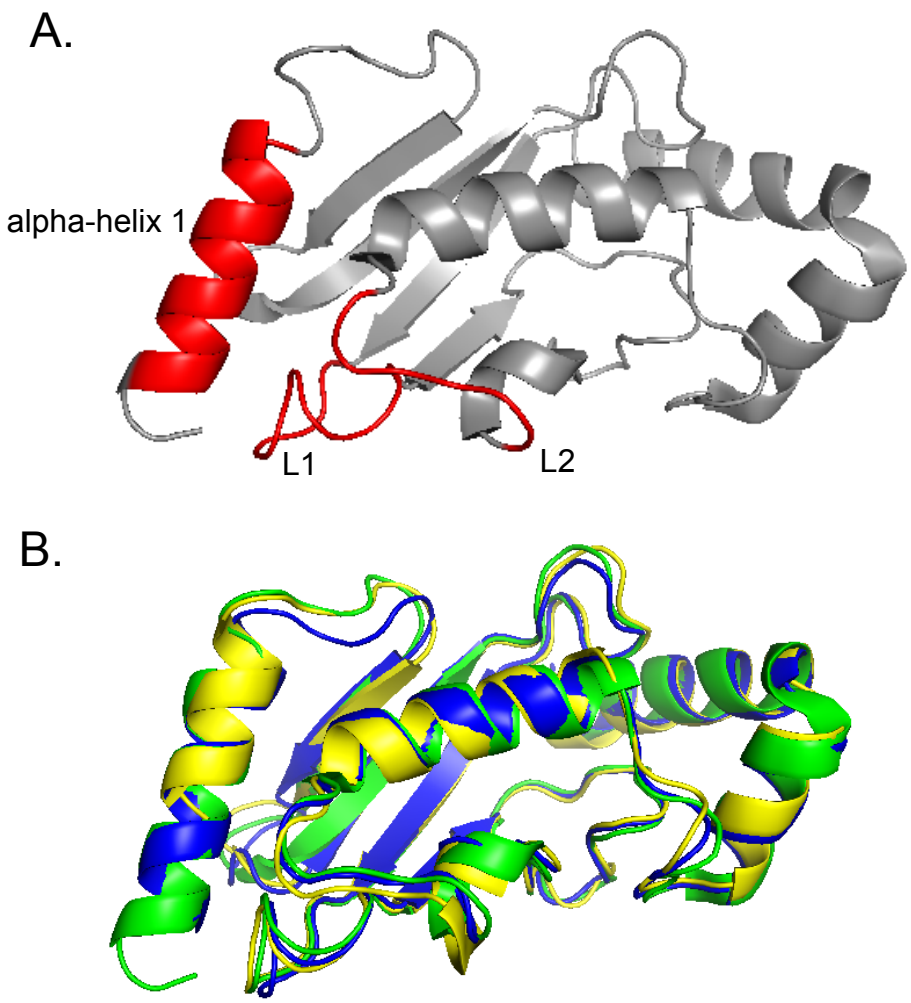


Figure 28. 3D structures of UBE2D1 protein. The crystallized structure of human UBE2D1 (PDB code: 2C4P.A) (Dodd and Read 2009) is shown with the catalytically active core (alpha-helix 1, L1 and L2 loops) highlighted in red (A). An overlay of the protein model prediction for elephant (yellow) and whale (blue) UBE2D1 onto the human crystallized structure (green) shows that the models match closely, but there are differences in the exact placement of loops and helices (B).

Because the amino acid changes occur in a functional domain of UBE2D1, we used protein structure prediction models to assess if the altered N-terminal amino acids in elephant and whale would be expected to alter the three-dimensional structure of the protein. We used a *de novo* secondary structure prediction model PredictProtein (Rost

and Liu 2003, Rost et al. 2004) to determine if the alpha-helix 1 would still fold given the different amino acid sequences of the elephant and whale relative to other mammals. This model predicted the same alpha-helix residues for the human, elephant and whale amino acid sequences. However, some amino acid changes were predicted to have a strong effect on the protein (Table 11) so we used ModWeb (Eswar et al. 2003) to model full tertiary structures of the protein sequences.

Amino Acid Change	Change in amino acid properties	Lineage(s) with change
M1L	NP; HPhb → NP; HPhb	whale & elephant
A2N	NP; HPhb → P; HPhl	elephant
A2S	NP; HPhb → P; HPhl	whale
L3I	NP; HPhb → NP; HPhb	whale & elephant
K4F	P(+); HPhl → NP; HPhb	whale & elephant
R5A	P(+); HPhl → NP; HPhb	whale & elephant
I6I	NP; HPhb → NP; HPhb	whale & elephant
Q7L	P; HPhl → NP; HPhb	whale & elephant
K8Q	P(+); HPhl → P; HPhl	whale & elephant

Table 11. N-terminal amino acid changes in UBE2D1 in elephant and whale. The first 8 residues of UBE2D1 and their changes seen in elephant and whale are shown. Residue changes that are predicted to have a strong (yellow) or weak effect (blue) are highlighted and those that are predicted to be neutral are white. Amino acid properties are abbreviated as NP: non-polar; P: polar; (+) positive charge; HPhb: hydrophobic; HPhl: hydrophilic.

ModWeb is a homology-based prediction tool that uses the available entries from the Protein Data Bank (PDB). The template chosen by the program to use for the whale model was a zebra fish UBE2D1 structure (PDB: 2OXQ) (Xu et al. 2008), and the template used for the elephant model was the human UBE2D2 (PDB: 4DDG.A) (Juang et

al. 2012). Unfortunately the user cannot control the model choice, and because the amino acid changes differ by one residue between elephant and whale, and neither matches the sequence of other mammalian UBE2D1, the chosen template models differ. The overall structure of the protein models predicted for the elephant and whale amino acid sequences closely match that of the human model; however there are slight changes in the exact positioning of loops and helices (Figure 28B). The elephant sequence resulted in a slightly truncated model, where the N-terminus does not include the full alpha-helix 1 because the first few amino acids are left out of the final model, which may be due to the PDB template choice made by ModWeb, or the A2N amino acid change may not result in a high confidence model of the full length alpha-helix.

ACCELERATED EVOLUTION IN TUMOR SUPPRESSOR GENES

Next we looked for evidence of accelerated evolution, which could suggest adaptation, or a change in selective pressure along a given lineage. We used codeML, which is part of the PAML software (Yang 2007) to search for tumor suppressor genes that may be evolving at a faster rate in elephants and/or whales. The same 501 filtered tumor suppressor genes that were used for the convergent evolution analysis were searched with PAML. We tested if the rates in the experimental model significantly differed from rates estimated by the null model in three different experiments: (1) the whale lineage independently, (2) the elephant lineage independently, and (3) both whale and elephant lineages at the same time to find any tumor suppressor genes that are undergoing accelerated evolution in both species.

In the first test, we found 16 proteins evolving at an accelerated rate specifically in the humpback whale (p-value < 0.05 after multiple testing correction by false discovery rate of q=0.05) (Table 12). Half of these genes (*HDAC2, YEATS4, DKK3, PLA2G16, DNAJA3, PERP, EIF2AK2, and CCNG1*) have *dN/dS* values greater than one, suggesting that they may be under positive selection. We used STRING (Franceschini et al. 2013) to look for how these genes may be related and we did not find any direct interactions or shared specific pathways. In the elephant genome we uncovered one protein sequence (*CDC73*) that is estimated to be evolving at an increased rate along the elephant branch and is statistically significant. Three genes are predicted to be evolving at an increased rate along both lineages, including *UBE2D1*, which we also found in our convergent evolution analysis (Table 12). All three shared genes showed evidence of relaxed purifying selection in elephant and whale (ω (other lineages) < ω (elephant and whale) < 1).

Because our results show many more genes for the whale compared to the elephant, we performed an additional test to make sure this was not an artifact of our *de novo* assembly. We used the same methods to test for accelerated evolution in the whale within a set of housekeeping genes, which had previously been used in a comparison of *dN/dS* to neuronal genes (Dorus, Vallender et al. 2004). No housekeeping genes were found to be evolving at a faster rate in the whale. This gives us confidence that the longer list of significant TSGs in whale compared to elephant is not likely caused by frequent sequencing errors.

Gene	Accelerated lineage	Accelerated ω	Non-accelerated ω	corrected P-value
HDAC2*	**whale**	**1.685**	**0.018**	**$2.7x10^{-10}$**
YEATS4*	**whale**	**8.1/0**	**0.011**	**$1.9x10^{-7}$**
PBRM1*	whale	0.235	0.026	0.001
PHF17*	whale	0.283	0.073	0.001
UCHL1	whale	0.683	0.047	0.002
YWHAQ	whale	0.734	0.021	0.002
SET*	whale	0.570	0.070	0.004
DKK3	**whale**	**1.452**	**0.175**	**0.005**
WDR11	whale	0.574	0.059	0.005
PLA2G16	**whale**	**7.7/0**	**0.144**	**0.006**
DNAJA3	**whale**	**5.1/0**	**0.057**	**0.006**
PCBP4	whale	0.368	0.018	0.009
PERP	**whale**	**3.649**	**0.182**	**0.013**
DAPK1	whale	0.501	0.036	0.024
EIF2AK2	**whale**	**10.341**	**0.518**	**0.025**
CCNG1	**whale**	**2.623**	**0.145**	**0.034**
CDC73*	elephant	0.308	0.005	0.021
UBE2D1	whale & elephant	0.598 & 0.309	0.000	0.000
CYLD	whale & elephant	0.151 & 0.159	0.036	0.014
PRDM2	whale & elephant	0.273 & 0.226	0.119	0.033

Table 12. Genes evolving at an accelerated rate. These genes were found to have higher *dN/dS* along the lineage(s) of interest (column 1). The p-value shown has been corrected for multiple testing by false discovery rate (q=0.05) and reflects that the model in which the specified lineage(s) had a different *dN/dS* fit the data better than the null model where each lineage has the same value. Those genes with a ratio greater than one are highlighted in bold. In cases where *dS* was estimated to be zero ω is shown as *dN/0.* An * next to the gene name indicates genes involved in chromatin organization and modification.

Though these accelerated genes have some commonalties because they are all involved in tumor suppression, we wanted to determine if there were any specific pathways or functions that stood out among this set of genes that we infer to be

undergoing accelerated evolution. We used DAVID (Huang da et al. 2009, Huang da et al. 2009) to investigate whether this accelerated gene set was enriched for any functions or pathways compared to the full list of tumor suppressor genes that we used in this study. Interestingly, we find that the genes evolving at an increased rate in the humpback whale and African elephant genomes are enriched for genes involved in chromatin organization and modification. Six genes (indicated in Table 12) are annotated with the terms "chromatin organization" and "chromosome organization" and the uncorrected p-values for the enrichment of these terms in the accelerated gene list over the background TSG list are 0.007 and 0.017, respectively. Other classifications that are significant include "histone modification", "covalent chromatin modification" and "chromatin modification" (Table 13). This suggests that regulation at the chromatin level may be important in the evolution of large, long-lived organisms, though the p-values are not significant ($p > 0.05$) after multiple testing correction.

Gene Ontology Term	**# of genes**	**uncorrected p-value**
Chromatin organization	6	0.007
Chromosome organization	6	0.017
Histone modification	4	0.019
Covalent chromatin modification	4	0.022
Chromatin modification	5	0.028

Table 13. Gene Ontology Terms Enriched in Accelerated Gene Set. This analysis was done with DAVID. The p-values shown are not corrected for multiple testing. After correction for multiple testing they are no longer less than 0.05.

METHODS

Gene Annotation for the Humpback Whale

Tumor suppresser genes from the Memorial Sloan Kettering CancerGenes database (Higgins, Claremont et al. 2007) were annotated with genBlastG (She et al. 2009, She, Chu et al. 2011) based on homology to the known human gene. The human amino acid sequences were chosen for gene prediction because it is a highly curated genome whereas the dolphin genome, which is currently the most closely related public genome to the humpback whale, is based solely on computational predictions and frequently has gaps in genes that are fully sequenced in the human.

We applied a number of filters to obtain a set of gene models for the whale orthologs of the human tumor suppressor genes. The top predicted gene sequence in the humpback whale genome had to cover at least 70% of the human query gene and be the reciprocal best BLAST hit. This is done to remove hits to closely related paralogous genes in cases where the true ortholog may not be assembled as well so it would not be the top hit. For example, if a gene is split between scaffolds, genBlastG will not use it to make a gene model unless one scaffold contains enough sequence to cover 70% of the human gene. This cutoff of 70% is a parameter set by the user and we chose 70% to match the requirements of CEGMA (Parra, Bradnam et al. 2009). From the original set of 830 tumor suppressor genes, we find that 724 are at least 70% coverage, but of these we only use 646, which are each the reciprocal best BLAST hit to the intended human protein. Additional filters were applied to this set for specific analyses as described below.

Copy Number of Tumor Suppressor Genes

We used the annotations from genBlastG to look for duplication events of tumor suppressor genes in the humpback whale genome. The program provides additional gene annotations for a given query gene if multiple locations are found within the target genome. For an alternative gene model to be considered an additional gene copy we required that the amino acid sequence from the gene model was also a reciprocal best BLAST hit to the human ortholog and that it could not overlap with the top gene model annotation. We only considered genes that had similar scoring gene models to the top ranked version (+/- 10 points) and shared similar sequence identity to the human ortholog as the top hit. The alternate gene model could have a better percent identity than the top rank hit; however if it was below the sequence identity of the best gene model for that protein, it had to be within 10%. These filters were used to make conservative calls for increased copy number of tumor suppressor genes in the whale genome. We further curated the resulting set of genes to determine if they were found to have a one-to-many relationship with the human gene in other mammals.

Obtaining Orthologous Gene Sequences

The gene list of 830 human tumor suppressor genes was downloaded from the CancerGenes database maintained by Memorial Sloan Kettering Cancer Center (Higgins, Claremont et al. 2007). A list of housekeeping genes used as a control were obtained from a study that had used this list as background to analyze accelerated evolution in

nervous system genes (Dorus, Vallender et al. 2004). Ortholog information for chimp (*Pan troglodytes*), dolphin (*Tursiops truncatus*), African elephant (*Loxodonta africana*), and hyrax (*Procavia capensis*) were downloaded from Ensembl BioMart (release 72). For the whale, we used the top prediction of orthologous coding sequence for each gene, which we annotated with genBlastG as described above. Only genes with a single gene copy in each organism were used for downstream analyses and the longest open reading frame was used in cases of alternative splicing. This restriction resulted in 501 tumor suppressor genes and 45 housekeeping genes.

Testing for Convergent Evolution

Each set of homologous gene coding sequences were translated to amino acids and aligned using Prank+F (Loytynoja and Goldman 2008). These alignments were used as input into PhyML (Guindon et al. 2010), optimizing for tree topology and branch length (parameters: -o tl –s SPR –d aa). The resulting tree topology was compared to a tree where elephant and whale were more closely related to each other than to any other species using the BioPerl module *Bio::Tree::Compatible* (Stajich et al. 2002). We manually inspected the resulting gene trees and alignments to remove those that were poor quality. To compare UBE2D1 sequences from other mammals, all mammalian sequences of UBE2D1 of length 147 (to avoid isoforms) were downloaded from Genbank and viewed in SeaView (Gouy et al. 2010). The protein models of UBE2D1 were viewed and overlaid using PyMol (Schrodinger 2013).

Testing for Accelerated Evolution

Each set of homologous gene sequences were aligned by codons using Prank+F (Loytynoja and Goldman 2008). The alignments were output in PHYLIP format and used as input into codeML, which is part of the PAML software package (Yang 2007). We ran codeML with the parameters specified in the control file, which can be found in the Appendix.

For each gene, codeML was run to test both the null model as well as the alternative hypothesis. The null model assumes that all branches have the same *dN/dS* (the ratio of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site). The alternative model tests whether a particular lineage has a different dN/dS value, indicating that is evolving at a different rate than the other lineages. We tested three alternative hypotheses for each gene:

1.) The African elephant lineage is evolving at a different rate than the others.

2.) The humpback whale lineage is evolving at a different rate than the others.

3.) The elephant and whale lineages are evolving at different rates than the others, but they are not necessarily evolving at the same rate as each other.

As suggested in the manual, codeML was run three times for each hypothesis tested for each gene (Yang 2007). This was done to check for reproducibility of the results since each instance begins with a random seed so the exact results are non-deterministic.

All *dN* and *dS* values were plotted in a histogram to detect extreme outliers, as poor quality alignments result in abnormally high values. Based on manual inspection of

the histogram, a cutoff of 0.048 for *dN* values and 0.31 for *dS* values was implemented for all downstream analyses. 95% of all estimated *dN* and *dS* values using codeML were less than or equal to these cutoffs. Values that were higher were often associated with poor alignments; therefore by implementing these cutoffs we helped to minimize the number of false positives caused by alignments.

The log-likelihood of the null and alternative models were output by codeML and used to compare the two hypotheses. We computed the log likelihood ratio test statistic D:

$$D = 2\times(H_1 - H_0),$$

where H_1 and H_0 are the log-likelihood values from the alternative and null models, respectively. The probability distribution of D is a chi-squared with degrees of freedom equal to :

$$df_D = df_1 - df_0,$$

where df_1 and df_0 are the degrees of freedom in the alternative and null models. To test for accelerated evolution along one lineage (elephant or whale), df_D is equal to 1. Additionally, to test both lineages at once, while allowing for different ω values along each, df_D equals 2. The values of *D* for the three repeated runs of codeML were averaged together. P-values were calculated based on the Chi-squared distribution with df_D degrees of freedom. The p-values were then corrected for multiple testing by false discovery rate (FDR) correction with a significance of 0.05 (i.e. q=0.05) (Benjamini and Hochberg 1995). We filtered for genes where ω along the lineage(s) of interest was

greater than estimated for the other branches. Alignments of genes with significant p-values after FDR correction were manually inspected and those with poor alignments were removed from the list.

DISCUSSION

We have used the humpback whale genome, which we *de novo* assembled, to reveal signs of selective pressures acting on tumor suppressor genes during the evolution of this large, long-lived mammal. We do not find any outliers in gene copy number of TSGs; however we do find two genes (*BECN1* and *TCEAL7*), which appear to have full gene duplications that include both the exons and introns. However, because this is a draft assembly, we cannot be sure that these copies are present in the full genome so future work would need to be done to address this uncertainty. We found that the second gene annotation for both *BECN1* and *TCELA7* genes on scaffolds that contain just that gene, which matches almost identically with the parent gene copy. Therefore, we cannot currently make primers to tease these apart, unless we can extend the sequence into unique regions for the shorter scaffold.

Our most surprising finding from the whale genome is the apparent convergent evolution of the amino acid sequence for the ubiquitin-conjugating enzyme (E2), *UBE2D1* (also known as *UbcH5a*). The process of adding ubiquitin onto a substrate typically requires the E1, E2 and E3 enzymes. The ubiquitin molecule is activated by the E1, transferred to the E2 and then the E3 transfers the ubiquitin to the substrate (Hershko et al. 1983). Depending on the position and the number of ubiquitin molecules on a

substrate, this process can mark a protein for degradation, change a protein's cellular location, or alter its activity (Welchman et al. 2005).

The predicted protein sequence of the humpback whale *UBE2D1* was annotated based on homology to the human sequence, not the elephant sequence, which should avoid bias in this result. We see perfect amino acid sequence conservation across other mammals, and because these changes are found within a catalytically active core domain which specifies the binding partners of the protein, we suspect that these differences in the elephant and whale would have an impact on the function of *UBE2D1*. Mutations to residues 5 and 9 in the N-terminal alpha-helix of UBE2D2 (an E2 in the same family as UBE2D1) expands its number of interacting E3 enzymes (van Wijk et al. 2009). We predict that the amino acid changes in the African elephant and humpback whale would also affect the number of interacting E3 ligases of UBE2D1, though we do not know if it would broaden the range of partners or create a specialist protein that has fewer binding partners, but perhaps with increased affinity for one another.

To reliably predict the binding partners of an E2 ubiquitin-conjugating enzyme with computational methods remains a challenge. Though there are still many unknowns surrounding these molecules and their interactions, a global yeast-two hybrid study revealed that there are at least 28 E3 binding partners of UBE2D1 (van Wijk, de Vries et al. 2009). UBE2D1, along with related E2s UBE2D2-4, act as hub proteins, so changing one would likely have a widespread impact on cellular functions (van Wijk, de Vries et al. 2009). One of the known binding partners of UBE2D1 is BRCA1, which functions as an E3 ligase when bound with BARD1 and UBE2D1 (Mallery et al. 2002). Disruption of

the interaction between BRCA1 and UBE2D1 has been associated with an increased risk of breast cancer (Morris et al. 2006) The E3 ubiquitin ligase activity of BRCA1 has been shown to ubiquitinate the sites of DNA double stranded breaks to signal for repair (Morris and Solomon 2004), which creates a logical connection to cancer incidence and ubiquitination. Additionally, polyubiquitination signals for protein degradation and can not only regulate processes like the cell cycle through this mechanism, but also functions to degrade abnormal proteins, which protects the cell from certain mutant phenotypes (Fredrickson and Gardner 2012).

Our analysis of accelerated evolution along the elephant and whale lineages suggests that regulation of cellular processes through chromatin remodeling is important in the evolution of large, long-lived organisms. We found genes involved in chromatin modification and organization are over-represented in the set of genes that we predicted to be evolving more rapidly in the humpback whale and African elephant genomes. Though most of these genes in the list are accelerated only in the whale genome, the one gene we find to be specific to the elephant is *CDC73,* which is involved in the methylation and mono-ubiquitination of histones (Hahn et al. 2012). Chromatin modifications can have a wide range of influences on a cell. For example, HDAC2 is thought to contribute to the deregulation of gene expression by aberrantly removing acetyl groups from histones and silencing tumor suppressor genes in malignant cells (Jung et al. 2012); however in normal cells it represses the transcription of many proto-oncogenes to regulate cell cycle progression (Ropero et al. 2008).

As in any genomic analysis, one caveat that we cannot avoid is that results are dependent on the quality of the genome assemblies and annotations. Looking for signatures of selective pressures depends highly on the multiple alignments of the genes being investigated. The quality of the alignments is largely influenced by the quality of the genome assembly and the annotations of the gene sequences. We enforced various filters to minimize the number of false positives due to poor alignments and then manually went through each gene that gave a significant result and discarded those that we felt were due to poor coding sequence annotation and/or alignment. This is an unfortunate bottleneck, but because we often rely on computational gene predictions for genome assemblies of unknown correctness, manually curating the results of evolutionary analyses remains necessary.

After carefully filtering all of our data, our results suggest that the evolution of cancer suppression in large, long-lived mammals is highly dependent on the precise regulation of gene and protein expression through chromatin remodeling and protein modifications, such as ubiquitination. We predict that this evolution is accomplished through the optimization of protein functions involved in these modifications, some of which we have revealed in this study. If the cellular levels of cancer-associated genes (i.e. proto-oncogenes and tumor suppressor genes) are not tightly regulated, cells can undergo transformation (Kitagawa et al. 2009). Previous work has also found that genes involved in the proteasome-ubiquitin pathway evolve faster along lineages that have increased longevity (Li and de Magalhaes 2013). Longevity is highly correlated with body size and in order to be large and live for many years, it is required that one can effectively suppress cancer.

CHAPTER 8: CONCLUSIONS AND FUTURE SUGGESTIONS

SCIENTIFIC CONTRIBUTIONS

Peto's paradox has remained an unsolved mystery for more than 30 years (Peto 1977). In this thesis, I present the first large-scale analysis of why cancer risk does not increase with body size and lifespan across species. I used hundreds of necropsy and death reports of captive mammals to provide empirical evidence that cancer incidence does not increase with body size and lifespan. I explored multiple computational models of cancer incidence in humans to quickly determine biologically feasible solutions to Peto's paradox which can serve as a focus of future experimental studies. I surveyed the copy number of cancer-associated genes in mammalian genomes and found that the African elephant genome contains 19 copies of the tumor suppressor gene *TP53*. Additionally, I analyzed the data collected by our collaborators at the University of Utah to demonstrate the novel result that African and Asian elephant cells undergo apoptosis at a much higher rate than human cells in response to γ-irradiation. The African elephant was previously the largest animal with an available genome assembly, so I undertook a *de novo* assembly of the humpback whale genome in order to expand our comparative genomic analyses. My analysis of the humpback whale genome revealed a set of tumor suppressor genes that are evolving at an accelerated rate and may play a role in increased cancer suppression.

One important impact of my work is to encourage more creative and innovative approaches to cancer research. Current therapies have harsh side effects and often do not cure the patient of the disease. It is important to focus more research efforts on improved prevention so we can decrease our lifetime risk of cancer and avoid these scenarios altogether (Etzioni, Urban et al. 2003). I believe that we can learn from the tumor suppression mechanisms that have evolved in large, long-lived animals to advance the field of cancer prevention. That said, the immediate translation of the results presented in this thesis to a clinical setting is not realistic. Further investigation of our comparative genomic findings in the African elephant and humpback whale is needed. Additionally, the field of comparative oncology and genomics should initiate new analyses and experiments to discover novel mechanisms of cancer suppression, which we may be able to mimic in humans as a new form of prevention.

SUGGESTIONS FOR THE FUTURE

If our current understanding of cancer is reasonable, there must be something fundamentally different in large, long-lived organisms to enhance their suppression of carcinogenesis. These mechanisms have allowed for the evolution of large bodies and extended lifespans without increasing the burden of cancer. In order to pursue this research and better out understanding of cancer suppression, we have had to make a number of assumptions and interpretations, which I will outline so that future work in this field can reassess these and determine the best course of action to move forward.

We have used mass as a proxy for cell number across species; however the composition of each animal is not identical. For example, whales have thick layers of blubber composed of adipose tissue and transformation of this tissue into malignant disease (e.g. liposarcoma) is rare (ACS 2013). Future work should consider whether it makes sense to compare cancer rates of individual tissue types instead of considering all cancers together. This would require accurate measurements of different organs across many species as well as the cancer incidence of the different tissues. We did not take this approach because cancer incidence in non-human animals is often not documented in a detailed manner so trying to make comparisons at this fine grain level would be challenging until this data improves.

Additionally there is the question of whether we should consider the lifetime risk of cancer or if it may make more sense to focus on pre-reproductive cancers when comparing rates across species. Selection acts to minimize pre-reproductive cancers (Graham 1992). Once an organism reaches an age at which are no longer able to reproduce, whether this is due to biological changes like menopause or sexual selection within the species, the need to suppress cancer decreases. The amount of time that an animal is able to reproduce and the time it takes to reach sexual maturity can vary not only between species but also between males and females of the same species. For example, female African elephant reach sexual maturity between 10-12 years of age and typically breed with a male shortly after. Though males also reach sexual maturity around the same time, the competition between males to mate with females strongly favors the largest males so most do not reproduce until 35 years old, with peak reproductive years from 45-53 (Holliseter-Smith et al. 2007). This suggests cancer suppression during this

time is still selected for in males. On the other hand, female African elephants begin to show a decrease in fertility around 35 years of age (Ward et al. 2009), so it would be interesting to see if there is a difference in the cancer rates of males and females. Cetaceans have a wide range of ages of sexual maturity and lifespans. Although most whales show no signs of reproductive senescence, the killer whale and the short-finned pilot whale go through changes similar to menopause in humans (Marsh 1986, Ward, Parsons et al. 2009). Because these species are unable to reproduce throughout their entire lives we would hypothesize that the selective pressure to suppress cancer would decrease with age in these animals relative to those that breed until old age. To account for these differences when comparing cancer incidence across species, future work should consider limiting the cancer rates to cases that occur only in early life and during reproductive years since selection may be acting differently in animals that have decline in fertility with age. Projects that expand on the work that I have outlined in the previous chapters should re-evaluate these assumptions about mass as a proxy for cell number and the proper comparisons for cancer rates in order to determine the best way to proceed with analyses.

Our analysis of computational models of cancer incidence provided us with the surprising result that a mere ~3-fold decrease in mutation rate can account for a one thousand-fold increase in body size (i.e. number of cells). Most researchers we have discussed with would have predicted that this value be significantly larger. The somatic mutation rate has not been measured for many species, especially non-model organisms like elephants and whales, so it would be interesting to get an estimate of these rates across animals spanning orders of magnitude in size and also ranging in lifespan. This

could be done *in vitro* using previously published techniques (Kondrashov 2003, Lynch 2010). Alternatively, we could make use methods that can estimate mutation rate from temporally spaced sequenced samples (Drummond et al. 2002). Longitudinal samples from the same individuals could be obtained from zoos to capture a wide range of species. It is not easy to get skin biopsies from zoo animals for research, and it would be challenging to sample the same location over time, which would confound the mutation rate estimates. However, blood draws are routinely done so I would suggest deep sequencing on DNA from blood in longitudinal samples to monitor mutations. Additionally, as the technology for genome sequencing from single cells improves, comparing the genomes of hundreds of single cells at each time point to the normal reference genome of that individual would provide more precise measures of the somatic mutation rate.

We have found 19 copies of the tumor suppressor gene *TP53* in the African elephant genome. Though 18 of these copies lack introns, we find evidence that they are actively transcribed. One caveat of this work is that we find additional copies of *TP53* that are expressed and we observe an increased apoptotic response to gamma-irradiation; however we have not been able to show a causal link between these two observations. There are many experiments that should be done to functionally characterize these genes and determine if they are responsible for the increased rate of apoptosis in response to gamma-irradiation. Two of the most interesting experiments, in my opinion, would be to knockdown the expression of the retrogenes in elephant cells with siRNA and to perform a knock-in of these genes in human cells. Both cell types would be subjected to the same irradiation experiment we discussed in Chapter 5, which would allow us to address

whether the expression of the *TP53* retrogenes is required for the high apoptotic response in elephant cells and if expressing these genes in human cells confers this hypersensitive phenotype. Additionally, we could test if the severity of the apoptotic response is dose dependent by doing partial knockdowns and also overexpressing the retrogenes. This could be coupled with genome-wide expression analyses to determine downstream targets or effectors of these genes. I would also recommend deep sequencing of the total RNA (with the exception of the ribosomal RNA), with technical and biological replicates, to find evidence of which specific retrogene loci are being expressed because currently we are unsure whether it is just a subset of them or all of them.

We were unable to produce protein from the retrogene clones using an *in vitro* coupled transcription/translation protocol, so it would be interesting to use mass spectrometry to determine if the retrogenes are producing protein. Additionally, the genomic location of these *TP53* copies is unknown, but is crucial information to have in order to understand how this gene was propagated within the genome and in what context each copy exists. We have obtained chromosome spreads for the African elephant with which we had hoped to do FISH; however we were stalled by finding the proper probe design. Because we only know approximately 2Kb of sequence for these retrogenes, and FISH targets are typically one to two orders of magnitude larger, it has been a challenge to find a way to create a probe that will produce enough fluorescence when it binds with this short of a sequence. However, the cycling-primed *in situ* labeling technique can use a short probe and amplifies the target sequence so that many copies then exist on the newly synthesized strand, which can then be detected by FISH (Talia et al. 2011). We can experiment with this technique in the future in addition to performing the experiments

proposed above to address the remaining questions surrounding the additional *TP53* copies in the African elephant. These experiments should be done for both the African and Asian elephants since we have preliminary work indicating that *TP53* retrogenes are also present in the Asian elephant and are currently cloning and sequencing them to determine their copy number and relationship to the copies found in the African elephant.

Our copy number analysis of tumor suppressor genes also revealed 8 annotated copies of *MAL* in the horse genome, and additionally two in the microbat genome. Unlike the retrogenes in the African elephant, these genes contain introns and appear as a tandem array on scaffold 15. It would be interesting to investigate these copies further to determine if they are in the genome and are not due to assembly error and if they are functional. Depending on the genomic validation, these genes could be expressed in human and mouse cells to deduce any enhanced mechanisms of cancer suppression that redundant *MAL* may confer.

The genomic analysis of the humpback whale provided many leads on interesting genes and pathways, which are promising directions for future investigation. Genes that appear duplicated should be verified; however because of the problems posed by the short scaffolds that the duplicates are on, which was discussed in Chapter 7, it will be important to further improve the genome assembly. The genome assembly is something that can always be updated and we should always strive for a complete genome. Long insert read libraries have proven to be very successful in other genomes (Li et al. 2010, Dong et al. 2013). Another avenue to pursue is optical mapping to create a more contiguous assembly. This technique was used for the domestic goat (*Capra hircus*)

genome and reduced the number of scaffolds to 315 from over 200,000 which increased the N50 from 3Mb to just over 16Mb (Dong, Xie et al. 2013). The minke whale genome was just released and though the assembly size (2.44Gb) is comparable to our *MegNov.v03* humpback whale assembly, they were able to obtain a significantly better scaffold N50 of 12.8Mb because they had four mate-pair libraries (2kb, 5kb, 10kb and 20kb) (Yim et al. 2013). The 10Kb and 20Kb libraries allow the reads to span large regions of tandem repeats and link contigs in non-repetitive regions. Interestingly, our *MegNov.v03* assembly has more nucleotide positions (A,C,G,T) even though they achieve a slightly longer assembly which means they have introduced more gaps, but have provided genomic context for the contigs. I believe if we could successfully prepare long insert mate-pair libraries, we too could get this contiguity with our genome assembly.

Further investigation of the *UBE2D1* gene and other genes we found to be undergoing accelerated evolution in the whale and elephant lineages could be undertaken even before an update of the genome assembly. The functional consequence of the predicted amino-acid changes in the alpha helix 1 of the elephant and whale UBE2D1 protein should be determined. This could be done with a yeast-two hybrid experiment to find if the E3 binding partners are affected by these changes (van Wijk, de Vries et al. 2009). It would also be interesting to introduce these amino acid changes in human cells and see if there is any phenotypic change in response to stresses, such as DNA damage. We also have collaborators who will collect fresh humpback whale blood when there is a live stranding, so we could use this in the same irradiation experiment that we did on the elephants to characterize the response to DNA damage caused by IR in whale PBMCs. If

these experiments yield interesting results that are suggestive of the increased tumor suppression of the UBE2D1 protein in elephant and whale, it would be interesting to pursue a mouse model in which the whale or elephant protein, containing the altered alpha helix 1 residues, is expressed instead of the endogenous copy in order to study possible changes in cancer susceptibility.

Due to computational constraints, we chose to run our analysis on a set of tumor suppressor genes that we annotated instead of doing a full genome annotation. In the future, when a full gene annotation is performed on the humpback whale genome, the accelerated evolution analysis can be repeated and one could look for enrichment of gene functions or pathway involvement in the accelerated gene set. This would be an unbiased approach and could result in not only genes that have been involved in the evolution of large bodies and long lifespans, but also in the transition from terrestrial life back to the marine environment and any other form of adaptation that a whale has undergone throughout its evolutionary history.

Though I have proposed multiple experiments to continue the research I have done in elephants and humpback whales, Peto's paradox transcends these two specific animals. Large bodies have evolved multiple times in the history of life, so each clade could have evolved different mechanism(s) to boost their tumor suppression abilities. Most of the hypotheses that have been proposed to explain Peto's paradox, as discussed in Chapter 1, have not been directly tested, and many related questions remain open.

As more genomes are sequenced, the power of comparative genomics will increase. An approach based on independent contrasts (Garland et al. 2005) of small and

large species within each clade could prove fruitful for identifying cancer suppression mechanisms. Large, long-lived organisms might have evolved to suppress cancer better than small animals by duplicating tumor suppressor genes (Nunney 1999, Leroi, Koufopanou et al. 2003) or eliminating some proto-oncogenes from the genome. A simple linear regression cannot be used to study whether a correlation exists between body size and the copy number of cancer related genes because this assumes independence of each genome. In reality, the genomes share many traits in common due to evolutionary descent from a common ancestor. An independent contrast model (Felsenstein 2003) should be used to partition the variance among species into comparisons that are independent of their evolutionary relationships. This analysis could be done by studying multiple clades, each composed of closely related species which have large variance in body size.

Marine mammals belonging to the order Cetacea are an ideal clade for this study since they range in size from small toothed whales like the harbor porpoise (~52.5Kg) (de Magalhães and Costa 2009) to the largest mammal on Earth, the blue whale (over 100,000 kg) (de Magalhães and Costa 2009). Though there is an extreme range in size, these species only diverged approximately 20-34 million years ago (Murphy, Pringle et al. 2007, Jackson, Baker et al. 2009). We have contributed to an effort to expand the number of sequenced cetacean genomes and there are groups around the world that are working on the bowhead whale, minke whale, blue whale, harbor porpoise and many others. Studies should focus on clades that include animals larger than humans, as opposed to looking at differences among various sized rodents or between mouse and

human, because the goal is to find a way of preventing cancer that is superior to endogenous tumor suppression mechanisms in humans.

Additionally, there are standard assays that could be used in comparative analyses to test many of the hypotheses for resolving Peto's paradox, including measurements of DNA damage repair (Olive and Banath 2006), somatic mutation rate (Drummond, Nicholls et al. 2002), telomere lengths (Canela et al. 2007), differentiation (Li et al. 2010) and proliferation (Woosley 1991, Minor 2008), apoptosis (Ribble et al. 2005), and reactive oxygen species (Afanasev 2009, Kundu et al. 2009).

Though this dissertation focused on cancer gene copy numbers and DNA-damage response, I also believe that basal metabolic rate plays a large role in the explanation of reduced cancer incidence in large animals. The metabolic rate impacts nearly all aspects of the cell and the byproducts, such as reactive oxygen species, are suspected to play a role in aging phenotypes and disease, including cancer (Ames 1989, Ku, Brunk et al. 1993, Stadtman 2004, Ivanova and Yankova 2013). Though our results suggest there are other mechanisms of cancer suppression in large, long-lived animals, they may merely supplement the cancer protection provided by a lower per cell metabolic rate.

We are among the first empirical endeavors into Peto's paradox and we hope to encourage other groups to pursue some of the research avenues I have suggested so that we may make advancements more rapidly. In order to truly change and innovate in cancer research we will need to address some shortcomings of the current state of the field. The majority of cancer research is done on a very small subset of organisms, which restricts our understanding of cancer to what we learn from those particular model

systems. Furthermore, the qualities of model organisms that make them ideal to work with in laboratory conditions (short lifespan and small body) are the very things that make them poor models for cancer suppression (Leroi, Koufopanou et al. 2003).

The lack of functional data for non-model organisms is also a major gap in the field. Function is often assumed from homology, which is not necessarily correct. For example, TSGs in *Drosophila* are largely non-overlapping with human tumor suppressors (Pearson and Sánchez Alvarado 2008). We make these assumptions in our own work because we currently only have computational gene predictions for many species, which is why follow-up studies should focus on the functional annotations of our genes of interest. We are also lacking robust epidemiological studies of cancer incidence in wildlife and captive populations. Captive populations will be useful for longitudinal studies and the predation-free environment will allow for better estimates of cancer rates. Studies that aim at a better understanding of the evolution of cancer suppression mechanisms will have to expand the variety of organisms that are studied in the laboratory setting and pursue both genomic and functional studies.

CONCLUSION

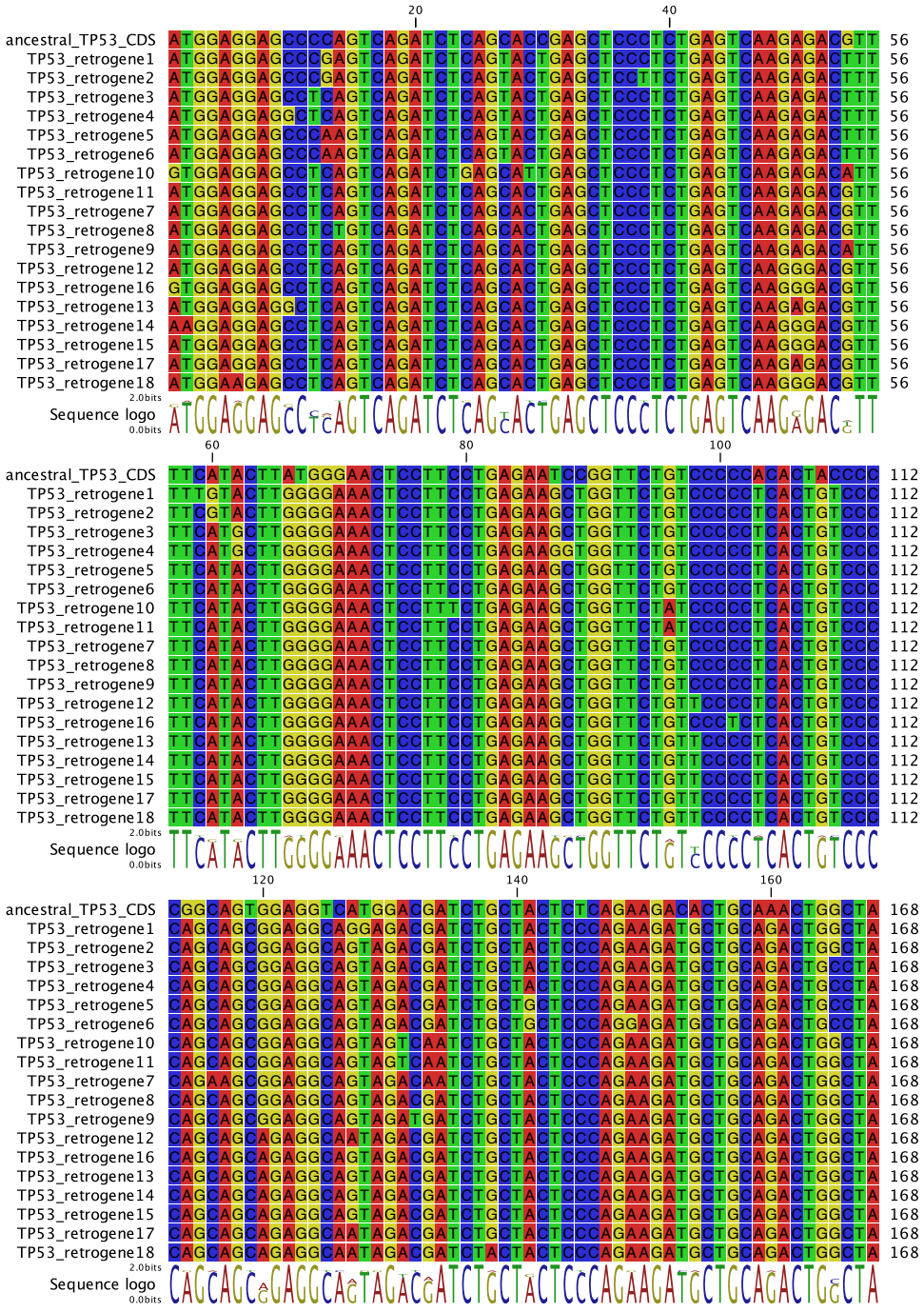
There has been no observed correlation between body size, longevity and lifetime cancer risk. Every additional cell and extra year of life should increase the probability of carcinogenesis. The fact that large, long-lived organisms are not over burdened by cancer suggests that they are more resistant to malignant transformation than smaller, more short-lived animals. Since large, long-lived organisms have achieved cancer suppression

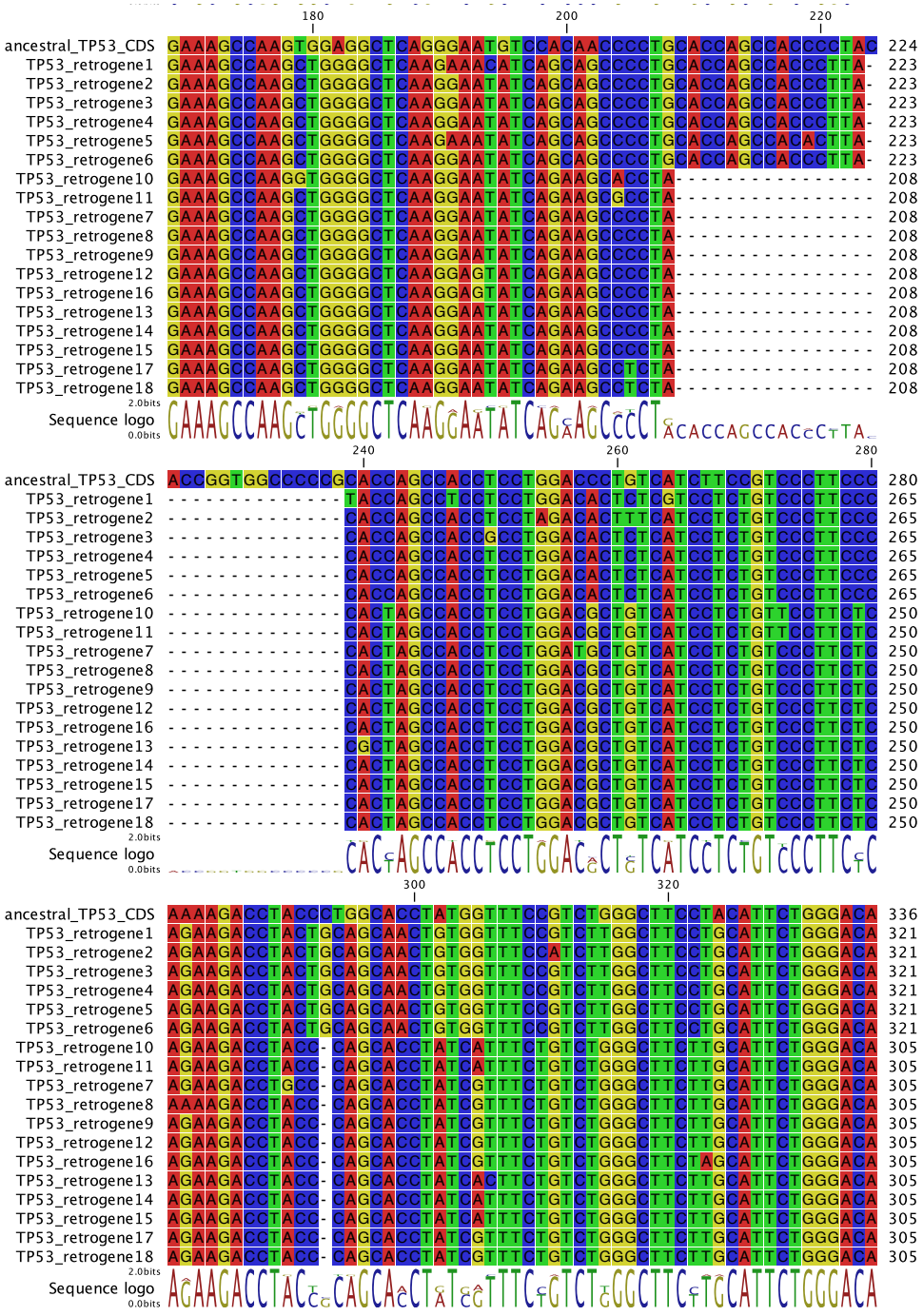
with minimal toxicity, this may be a fruitful avenue for cancer prevention research. People have only been invested in cancer research for decades while evolution has been tuning cancer suppression mechanisms for over a billion years. If we can harness the cancer suppression mechanisms of large, long-lived organisms, then we could potentially eradicate cancer as a public health threat in humans.

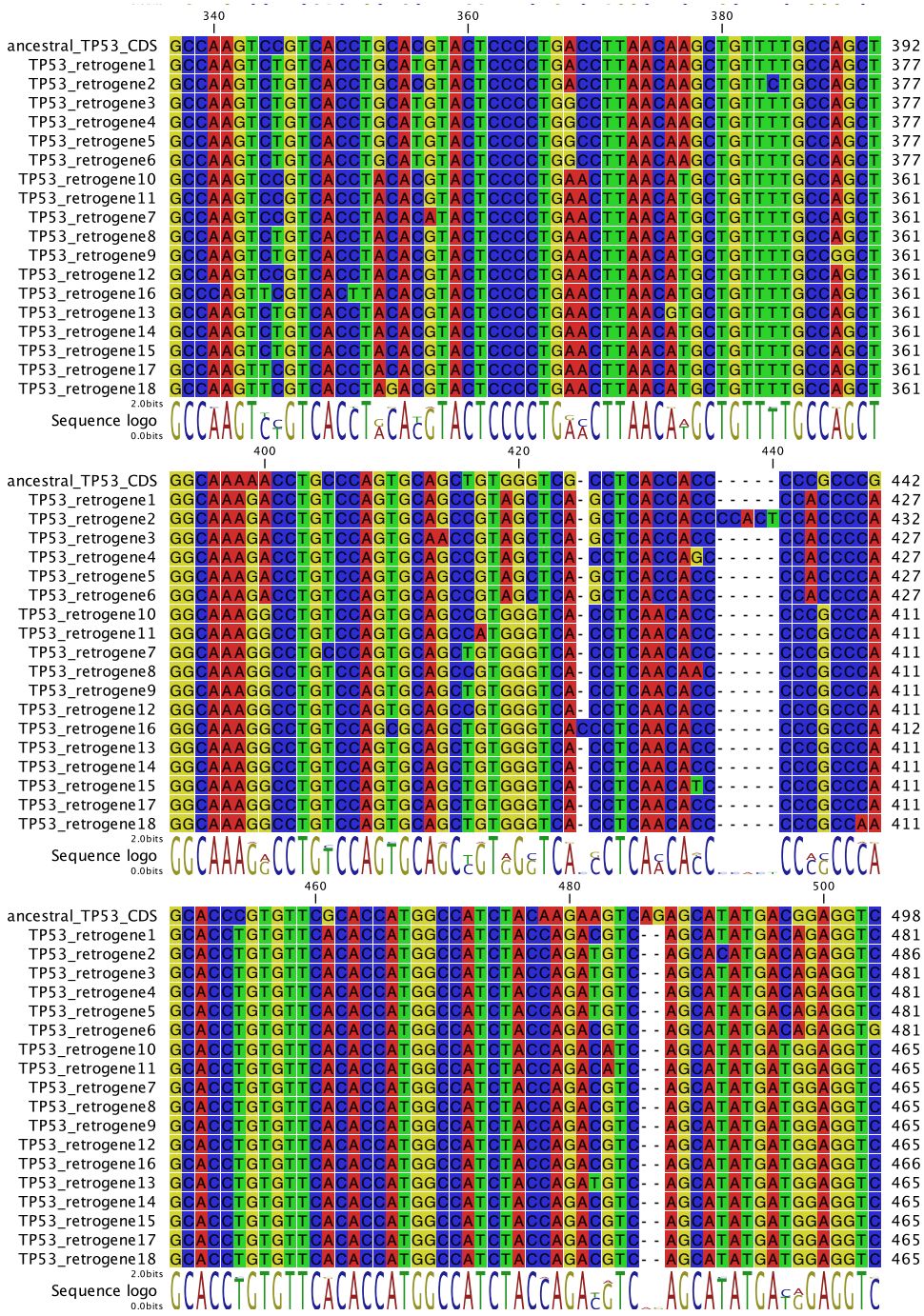
APPENDIX

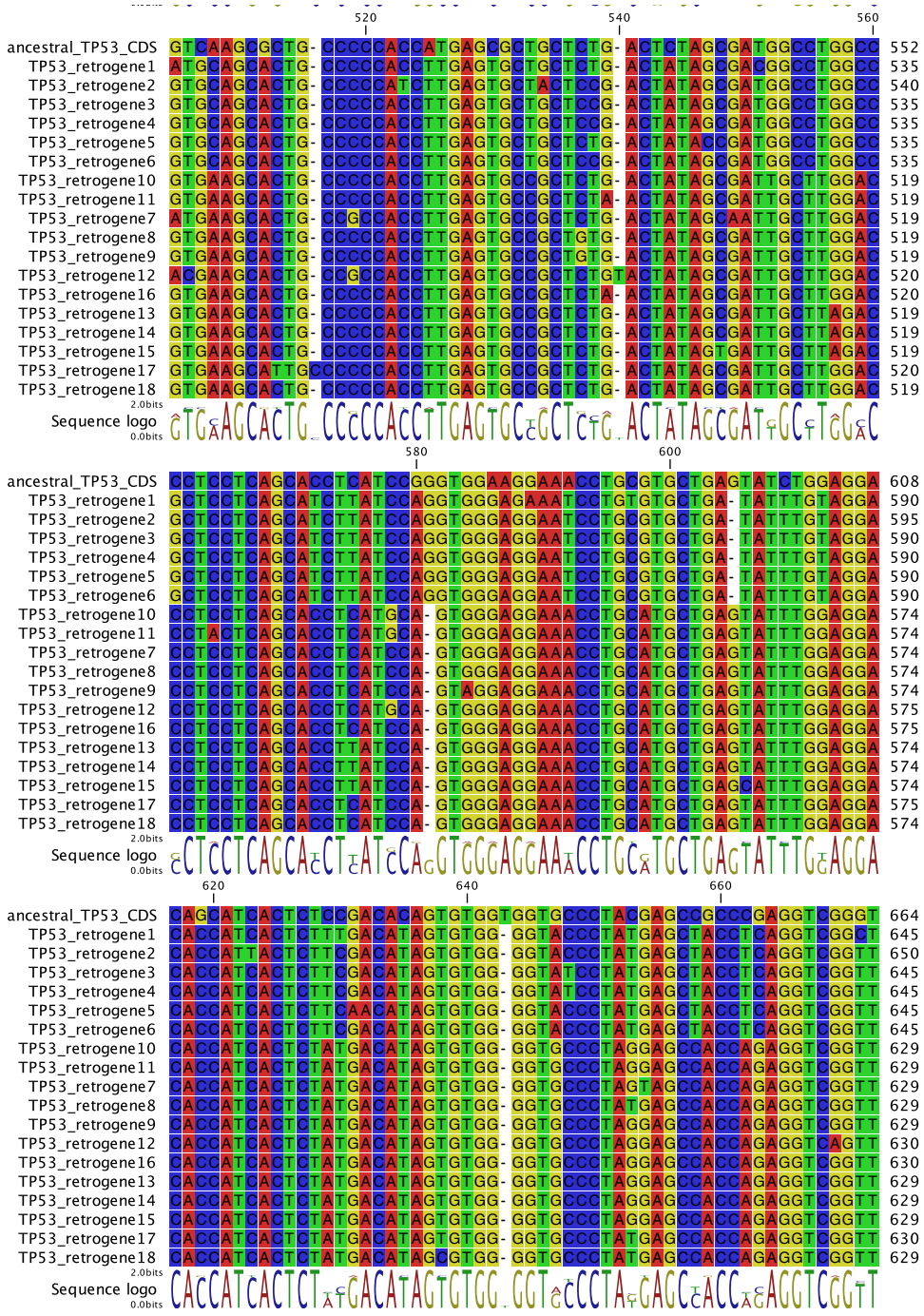
MULTIPLE ALIGNMENT OF *TP53* RETROGENES

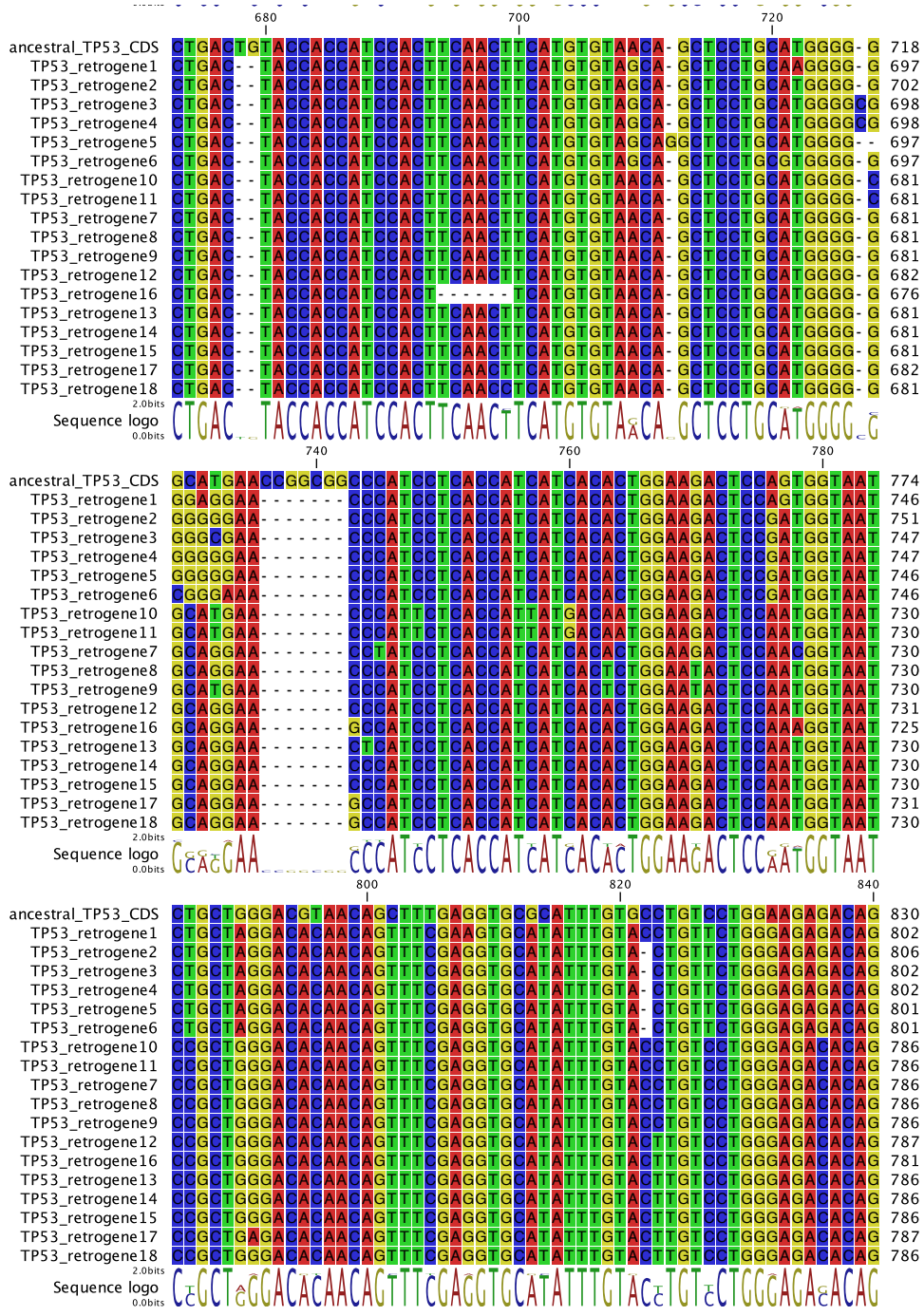
The below alignment was generated with MUSCLE using CLC Main Workbench 7 and shows the *TP53* retrogenes aligned to the coding sequence of the ancestral copy.

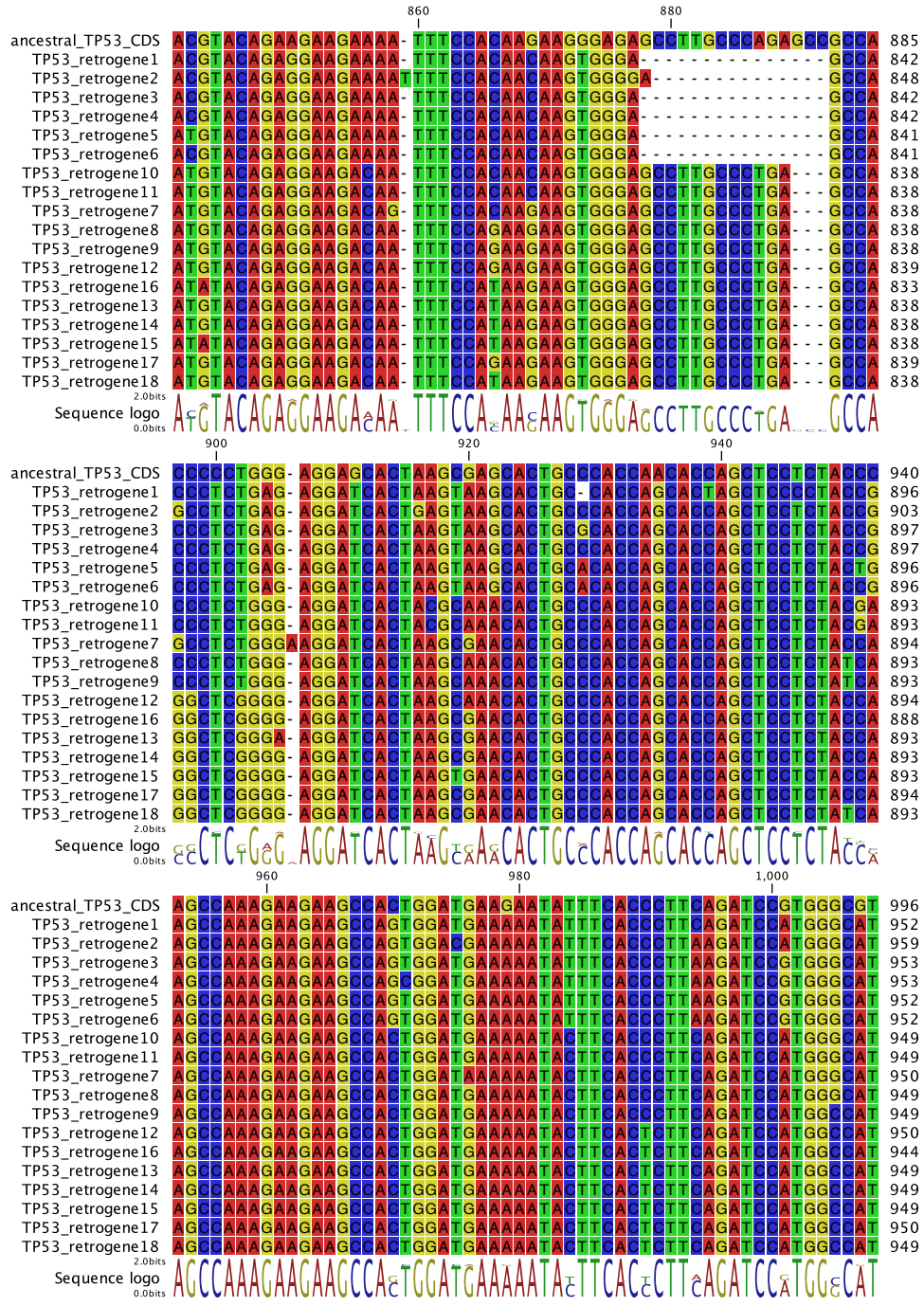


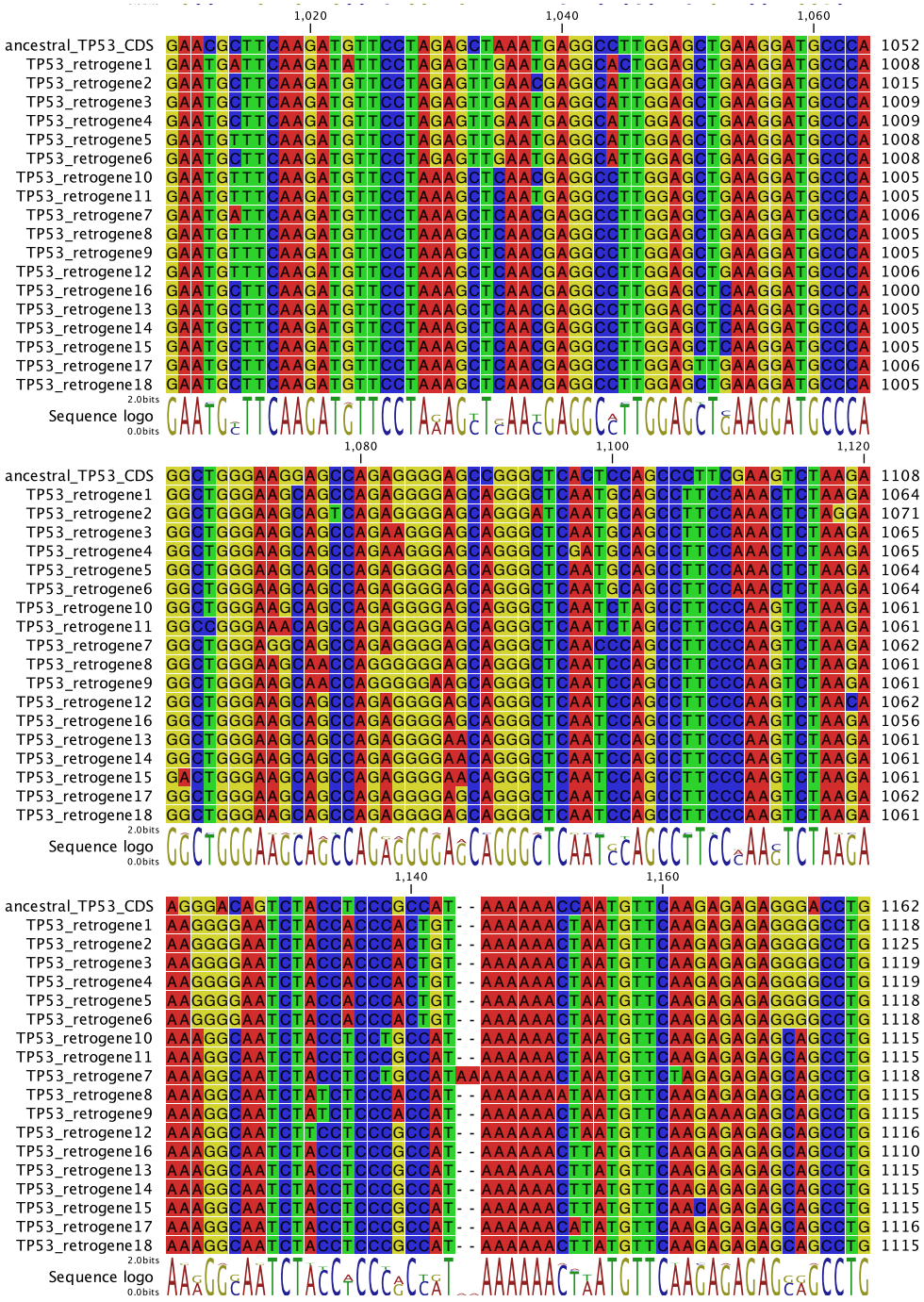

180
200
220
ancestral_TP53_CDS
TP53_retrogene1
TP53_retrogene2
TP53_retrogene3
TP53_retrogene4
TP53_retrogene5
TP53_retrogene6
TP53_retrogene10
TP53_retrogene11
TP53_retrogene7
TP53_retrogene8
TP53_retrogene9
TP53_retrogene12
TP53_retrogene16
TP53_retrogene13
TP53_retrogene14
TP53_retrogene15
TP53_retrogene17
TP53_retrogene18
2.0bits
Sequence logo
0.0bits
240
260
280
300
320

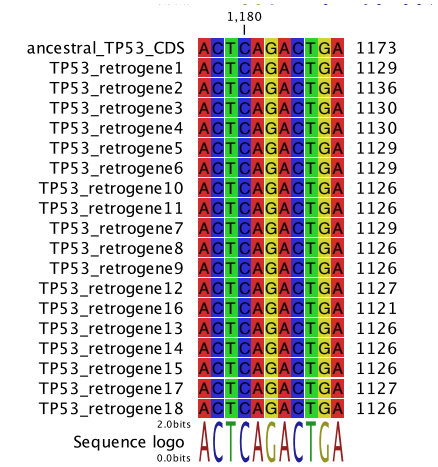

340
360
380
ancestral_TP53_CDS 392
TP53_retrogene1 377
TP53_retrogene2 377
TP53_retrogene3 377
TP53_retrogene4 377
TP53_retrogene5 377
TP53_retrogene6 377
TP53_retrogene10 361
TP53_retrogene11 361
TP53_retrogene7 361
TP53_retrogene8 361
TP53_retrogene9 361
TP53_retrogene12 361
TP53_retrogene16 361
TP53_retrogene13 361
TP53_retrogene14 361
TP53_retrogene15 361
TP53_retrogene17 361
TP53_retrogene18 361
2.0bits
Sequence logo
0.0bits
400
420
440
ancestral_TP53_CDS 442
TP53_retrogene1 427
TP53_retrogene2 432
TP53_retrogene3 427
TP53_retrogene4 427
TP53_retrogene5 427
TP53_retrogene6 427
TP53_retrogene10 411
TP53_retrogene11 411
TP53_retrogene7 411
TP53_retrogene8 411
TP53_retrogene9 411
TP53_retrogene12 411
TP53_retrogene16 412
TP53_retrogene13 411
TP53_retrogene14 411
TP53_retrogene15 411
TP53_retrogene17 411
TP53_retrogene18 411
2.0bits
Sequence logo
0.0bits
460
480
500
ancestral_TP53_CDS 498
TP53_retrogene1 481
TP53_retrogene2 486
TP53_retrogene3 481
TP53_retrogene4 481
TP53_retrogene5 481
TP53_retrogene6 481
TP53_retrogene10 465
TP53_retrogene11 465
TP53_retrogene7 465
TP53_retrogene8 465
TP53_retrogene9 465
TP53_retrogene12 465
TP53_retrogene16 466
TP53_retrogene13 465
TP53_retrogene14 465
TP53_retrogene15 465
TP53_retrogene17 465
TP53_retrogene18 465
2.0bits
Sequence logo
0.0bits


520
540
560
ancestral_TP53_CDS 552
TP53_retrogene1 535
TP53_retrogene2 540
TP53_retrogene3 535
TP53_retrogene4 535
TP53_retrogene5 535
TP53_retrogene6 535
TP53_retrogene10 519
TP53_retrogene11 519
TP53_retrogene7 519
TP53_retrogene8 519
TP53_retrogene9 519
TP53_retrogene12 520
TP53_retrogene16 520
TP53_retrogene13 519
TP53_retrogene14 519
TP53_retrogene15 519
TP53_retrogene17 520
TP53_retrogene18 519
2.0bits
Sequence logo
0.0bits
580
600
ancestral_TP53_CDS 608
TP53_retrogene1 590
TP53_retrogene2 595
TP53_retrogene3 590
TP53_retrogene4 590
TP53_retrogene5 590
TP53_retrogene6 590
TP53_retrogene10 574
TP53_retrogene11 574
TP53_retrogene7 574
TP53_retrogene8 574
TP53_retrogene9 574
TP53_retrogene12 575
TP53_retrogene16 575
TP53_retrogene13 574
TP53_retrogene14 574
TP53_retrogene15 574
TP53_retrogene17 575
TP53_retrogene18 574
2.0bits
Sequence logo
0.0bits
620
640
660
ancestral_TP53_CDS 664
TP53_retrogene1 645
TP53_retrogene2 650
TP53_retrogene3 645
TP53_retrogene4 645
TP53_retrogene5 645
TP53_retrogene6 645
TP53_retrogene10 629
TP53_retrogene11 629
TP53_retrogene7 629
TP53_retrogene8 629
TP53_retrogene9 629
TP53_retrogene12 630
TP53_retrogene16 630
TP53_retrogene13 629
TP53_retrogene14 629
TP53_retrogene15 629
TP53_retrogene17 630
TP53_retrogene18 629
2.0bits
Sequence logo
0.0bits


680
700
720
ancestral_TP53_CDS
TP53_retrogene1
TP53_retrogene2
TP53_retrogene3
TP53_retrogene4
TP53_retrogene5
TP53_retrogene6
TP53_retrogene10
TP53_retrogene11
TP53_retrogene7
TP53_retrogene8
TP53_retrogene9
TP53_retrogene12
TP53_retrogene16
TP53_retrogene13
TP53_retrogene14
TP53_retrogene15
TP53_retrogene17
TP53_retrogene18
2.0bits
Sequence logo
0.0bits
740
760
780
800
820
840


860
880
ancestral_TP53_CDS
TP53_retrogene1
TP53_retrogene2
TP53_retrogene3
TP53_retrogene4
TP53_retrogene5
TP53_retrogene6
TP53_retrogene10
TP53_retrogene11
TP53_retrogene7
TP53_retrogene8
TP53_retrogene9
TP53_retrogene12
TP53_retrogene16
TP53_retrogene13
TP53_retrogene14
TP53_retrogene15
TP53_retrogene17
TP53_retrogene18
2.0bits
Sequence logo
0.0bits
900
920
940
960
980
1,000


1,020
1,040
1,060
ancestral_TP53_CDS 1052
TP53_retrogene1 1008
TP53_retrogene2 1015
TP53_retrogene3 1009
TP53_retrogene4 1009
TP53_retrogene5 1008
TP53_retrogene6 1008
TP53_retrogene10 1005
TP53_retrogene11 1005
TP53_retrogene7 1006
TP53_retrogene8 1005
TP53_retrogene9 1005
TP53_retrogene12 1006
TP53_retrogene16 1000
TP53_retrogene13 1005
TP53_retrogene14 1005
TP53_retrogene15 1005
TP53_retrogene17 1006
TP53_retrogene18 1005
2.0bits
Sequence logo
0.0bits
1,080
1,100
1,120
ancestral_TP53_CDS 1108
TP53_retrogene1 1064
TP53_retrogene2 1071
TP53_retrogene3 1065
TP53_retrogene4 1065
TP53_retrogene5 1064
TP53_retrogene6 1064
TP53_retrogene10 1061
TP53_retrogene11 1061
TP53_retrogene7 1062
TP53_retrogene8 1061
TP53_retrogene9 1061
TP53_retrogene12 1062
TP53_retrogene16 1056
TP53_retrogene13 1061
TP53_retrogene14 1061
TP53_retrogene15 1061
TP53_retrogene17 1062
TP53_retrogene18 1061
2.0bits
Sequence logo
0.0bits
1,140
1,160
ancestral_TP53_CDS 1162
TP53_retrogene1 1118
TP53_retrogene2 1125
TP53_retrogene3 1119
TP53_retrogene4 1119
TP53_retrogene5 1118
TP53_retrogene6 1118
TP53_retrogene10 1115
TP53_retrogene11 1115
TP53_retrogene7 1118
TP53_retrogene8 1115
TP53_retrogene9 1115
TP53_retrogene12 1116
TP53_retrogene16 1110
TP53_retrogene13 1115
TP53_retrogene14 1115
TP53_retrogene15 1115
TP53_retrogene17 1116
TP53_retrogene18 1115
2.0bits
Sequence logo
0.0bits



SGA PIPELINE FOR FOSMID ASSEMBLY

```
CPU=8
DISTANCE_EST=DistanceEst

$SGA preprocess -p 1 -o SGA_pp_reads.fastq -v $1 $2
$SGA index -t 8 --no-reverse SGA_pp_reads.fastq
$SGA correct -v -k 31 --metrics=kmer_errors.log -o reads.ec.fastq --learn -t 8
SGA_pp_reads.fastq

MIN_OVERLAP=35
ASSEMBLE_OVERLAP=40

TRIM_LENGTH=200
MIN_CONTIG_LENGTH=1000
MIN_PAIRS=3
D=2000000

#
# Primary (contig) assembly
#
# Index the corrected data.
$SGA index -d $D -t $CD=2000000 -t $CPU reads.ec.fastq
```

```
# Remove exact-match duplicates and reads with low-frequency k-mers
$SGA filter -x 2 -t $CPU reads.ec.fastq

# Compute the structure of the string graph
$SGA overlap -m $MIN_OVERLAP -t $CPU reads.ec.filter.pass.fa

# Perform the contig assembly
$SGA assemble -m $ASSEMBLE_OVERLAP --min-branch-length $TRIM_LENGTH -
o primary reads.ec.filter.pass.asqg.gz

#
# Scaffolding
#

IN1=$1
IN2=$2
PRIMARY_CONTIGS=primary-contigs.fa
PRIMARY_GRAPH=primary-graph.asqg.gz

# Align the reads to the contigs
$BWA_BIN index $PRIMARY_CONTIGS
$BWA_BIN aln -t $CPU $PRIMARY_CONTIGS $IN1 > $IN1.sai
$BWA_BIN aln -t $CPU $PRIMARY_CONTIGS $IN2 > $IN2.sai
$BWA_BIN sampe $PRIMARY_CONTIGS $IN1.sai $IN2.sai $IN1 $IN2 |
$SAMTOOLS_BIN view -Sb - > libPE.bam

# Convert the BAM file into a set of contig-contig distance estimates
$BAM2DE_BIN -n $MIN_PAIRS -m $MIN_CONTIG_LENGTH --prefix libPE
libPE.bam

# Compute copy number estimates of the contigs
$ASTAT_BIN -m $MIN_CONTIG_LENGTH libPE.bam > libPE.astat


# Build the scaffolds
$SGA scaffold -m $MIN_CONTIG_LENGTH -a libPE.astat -o scaffolds.scaf --pe
libPE.de $PRIMARY_CONTIGS

# Convert the scaffolds to FASTA format
$SGA scaffold2fasta --use-overlap --write-unplaced -m $MIN_CONTIG_LENGTH -a
$PRIMARY_GRAPH -o sga-scaffolds.fa scaffolds.scaf
```

MaSuRCA Parameters

GRAPH_KMER_SIZE=auto
CA_PARAMETERS = ovlMerSize=30 cgwErrorRate=0.15 ovlMemory=4GB
WINDOW=10
MAX_ERR_PER_WINDOW=3
TRIM_PARAM=2
EXTEND_JUMP_READS=0
NUM_THREADS= 32
JF_SIZE=30000000000
DO_HOMOPOLYMER_TRIM=0
USE_LINKING_MATES=1

PAML Control Files

Null Model Ctl File:

seqfile = <alignment.phy>
treefile = <null.model.tree>
outfile = <null.paml.out>
model = 0
noisy = 0
verbose = 0
runmode = 0
seqtype = 1
CodonFreq = 3
aaDist = 0
NSsites = 0
icode = 0
Mgene = 0
fix_kappa = 0
kappa = 2
fix_omega = 0
omega = .5
fix_alpha = 1
alpha = 0
Malpha = 0
ncatG = 1
clock = 0
getSE = 0
RateAncestor = 0
cleandata = 1

Experimental Model Ctl File:

seqfile = <alignment.phy>
treefile = <exp.model.tree>
outfile = <exp.paml.out>
model = 2
noisy = 0
verbose = 0
runmode = 0
seqtype = 1
CodonFreq = 3
aaDist = 0
NSsites = 0
icode = 0
Mgene = 0
fix_kappa = 0
kappa = 2
fix_omega = 0
omega = .5
fix_alpha = 1
alpha = 0
Malpha = 0
ncatG = 1
clock = 0
getSE = 0
RateAncestor = 0
cleandata = 1

REFERENCES

ACS (2013). American Cancer Society. Cancer Facts & FIgures 2013.

Adamson, E. D. (1987). "Oncogenes in development." Development **99**(4): 449-471.

Adegoke, J. A., U. Arnason and B. Widegren (1993). "Sequence organization and evolution, in all extant whalebone whales, of a DNA satellite with terminal chromosome localization." Chromosoma **102**(6): 382-388.

Adelman, R., R. L. Saul and B. N. Ames (1988). "Oxidative damage to DNA: relation to species metabolic rate and life span." Proc Natl Acad Sci U S A **85**(8): 2706-2708.

Afanasev, I. (2009). "Detection of superoxide in cells, tissues and whole organisms." Front Biosci (Elite Ed) **1**: 153-160.

Albanes, D. (1998). "Height, early energy intake, and cancer. Evidence mounts for the relation of energy intake to adult malignancies." BMJ (Clinical research ed.) **317**(7169): 1331-1332.

Alonso, M. A. and S. M. Weissman (1987). "cDNA cloning and sequence of MAL, a hydrophobic protein associated with human T-cell differentiation." Proc Natl Acad Sci U S A **84**(7): 1997-2001.

Altman, A. J. and A. D. Schwartz (1978). "Malignant diseases of infancy, childhood and adolescence." Major Probl Clin Pediatr **18**: 1-515.

Ames, B. N. (1989). "Endogenous oxidative DNA damage, aging, and cancer." Free Radic Res Commun **7**(3-6): 121-128.

Andervont, H. B. and T. B. Dunn (1962). "Occurrence of tumors in wild house mice." J Natl Cancer Inst **28**: 1153-1163.

Arlt, A. and H. Schafer (2002). "NFkappaB-dependent chemoresistance in solid tumors." Int J Clin Pharmacol Ther **40**(8): 336-347.

Arnason, U., M. Hoglund and B. Widegren (1984). "Conservation of highly repetitive DNA in cetaceans." Chromosoma **89**(3): 238-242.

Arnason, U., I. F. Purdom and K. W. Jones (1978). "Conservation and chromosomal localization of DNA satellites in balenopterid whales." Chromosoma **66**(2): 141-159.

Arnason, U. and B. Widegren (1984). "Different rates of divergence in highly repetitive DNA of cetaceans." Hereditas **101**(2): 171-177.

Arnason, U. and B. Widegren (1989). "Composition and chromosomal localization of cetacean highly repetitive DNA with special reference to the blue whale, Balaenoptera musculus." Chromosoma **98**(5): 323-329.

Aronesty, E. (2011). ea-utils: Command-line tools for processing biological sequence data. http://code.google.com/p/ea-utils.

Axelrod, R., D. E. Axelrod and K. J. Pienta (2006). "Evolution of cooperation among tumor cells." Proc Natl Acad Sci U S A **103**(36): 13474-13479.

Baker, C. S., A. Perry, J. L. Bannister, et al. (1993). "Abundant mitochondrial DNA variation and world-wide population structure in humpback whales." Proc Natl Acad Sci U S A **90**(17): 8239-8243.

Beerenwinkel, N., T. Antal, D. Dingli, et al. (2007). "Genetic progression and the waiting time to cancer." PLoS Comput Biol **3**(11): e225.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society **57**(1): 289-300.
Bernier, S. G., N. Taghizadeh, C. D. Thompson, et al. (2005). "Methionine aminopeptidases type I and type II are essential to control cell proliferation." J Cell Biochem **95**(6): 1191-1203.
Bernstein, C., H. Bernstein, C. M. Payne, et al. (2002). "DNA repair/pro-apoptotic dual-role proteins in five major DNA repair pathways: fail-safe protection against carcinogenesis." Mutat Res **511**(2): 145-178.
Bielas, J. H., K. R. Loeb, B. P. Rubin, et al. (2006). "Human cancers express a mutator phenotype." Proc Natl Acad Sci U S A **103**(48): 18238-18242.
Bininda-Emonds, O. R., M. Cardillo, K. E. Jones, et al. (2007). "The delayed rise of present-day mammals." Nature **446**(7135): 507-512.
Bradnam, K. R., J. N. Fass, A. Alexandrov, et al. (2013). "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species." Gigascience **2**(1): 10.
Bu, D., K. Yu, S. Sun, et al. (2012). "NONCODE v3.0: integrative annotation of long noncoding RNAs." Nucleic Acids Res **40**(Database issue): D210-215.
Buffenstein, R. (2005). "The naked mole-rat: a new long-living model for human aging research." The journals of gerontology. Series A, Biological sciences and medical sciences **60**(11): 1369-1377.
Buffenstein, R. and J. U. Jarvis (2002). "The naked mole rat--a new record for the oldest living rodent." Sci Aging Knowledge Environ **2002**(21): pe7.
Cairns, J. (1975). "Mutation selection and the natural history of cancer." Nature **255**(5505): 197-200.
Calabrese, P. and D. Shibata (2010). "A simple algebraic cancer equation: calculating how cancers may arise with normal mutation rates." BMC cancer **10**(1): 3.
Calabrese, P., S. Tavare and D. Shibata (2004). "Pretumor progression: clonal evolution of human stem cell populations." The American journal of pathology **164**(4): 1337-1346.
Campisi, J. (1997). "Aging and cancer: the double-edged sword of replicative senescence." J Am Geriatr Soc **45**(4): 482-488.
Campisi, J. (2001). "Cellular senescence as a tumor-suppressor mechanism." Trends in cell biology **11**(11): S27-31.
Campisi, J. (2003). "Cancer and ageing: rival demons?" Nat Rev Cancer **3**(5): 339-349.
Campisi, J. (2005). "Senescent cells, tumor suppression, and organismal aging: good citizens, bad neighbors." Cell **120**(4): 513-522.
Canela, A., P. Klatt and M. A. Blasco (2007). "Telomere length analysis." Methods Mol Biol **371**: 45-72.
Cao, W., Z. Y. Zhang, Q. Xu, et al. (2010). "Epigenetic silencing of MAL, a putative tumor suppressor gene, can contribute to human epithelium cell carcinoma." Mol Cancer **9**: 296.
Caulin, A. F. and C. C. Maley (2011). "Peto's Paradox: evolution's prescription for cancer prevention." Trends in ecology & evolution **26**(4): 175-182.
Chen, F. C. and W. H. Li (2001). "Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees." Am J Hum Genet **68**(2): 444-456.
Cheong, K. H., D. Zacchetti, E. E. Schneeberger, et al. (1999). "VIP17/MAL, a lipid raft-associated protein, is involved in apical transport in MDCK cells." Proc Natl Acad Sci U S A **96**(11): 6241-6248.

Chu, E. H., M. Boehnke, S. M. Hanash, et al. (1988). "Estimation of mutation rates based on the analysis of polypeptide constituents of cultured human lymphoblastoid cells." Genetics **119**(3): 693-703.
Clapham, P., S. Young and J. R. Brownell (1999). "Baleen Whales: Conservation Issues and the Status of the Most Endangered Populations." Mammal Review(29): 35-60.
Clevers, H. (2005). "Stem cells, asymmetric division and cancer." Nat Genet **37**(10): 1027-1028.
d'Adda di Fagagna, F., P. M. Reaper, L. Clay-Farrace, et al. (2003). "A DNA damage checkpoint response in telomere-initiated senescence." Nature **426**(6963): 194-198.
Danilov, A., M. Shaposhnikov, E. Plyusnina, et al. (2013). Selective anticancer agents suppress aging in Drosophila.
Davenport, M. P., R. L. Ward and N. J. Hawkins (2002). "The null oncogene hypothesis and protection from cancer." J Med Genet **39**(1): 12-14.
Dawe, C. J., J. C. Harshbarger and Smithsonian Institution. (1969). Neoplasms and related disorders of invertebrate and lower vertebrate animals; [proceedings. Bethesda, Md.,, U.S. National Cancer Institute; for sale by the Supt. of Docs.
de Magalhães, J. P. and J. Costa (2009). "A database of vertebrate longevity records and their relation to other life-history traits." Journal of evolutionary biology **22**(8): 1770-1774.
DeGregori, J. (2011). "Evolved tumor suppression: why are we so good at not getting cancer?" Cancer Res **71**(11): 3739-3744.
Deng, W. G., H. Kawashima, G. Wu, et al. (2007). "Synergistic tumor suppression by coexpression of FUS1 and p53 is associated with down-regulation of murine double minute-2 and activation of the apoptotic protease-activating factor 1-dependent apoptotic pathway in human non-small cell lung cancer cells." Cancer Res **67**(2): 709-717.
Do, B. H., A. S. Wu, J. Maley, et al. (2013). "Automatic retrieval of bone fracture knowledge using natural language processing." J Digit Imaging **26**(4): 709-713.
Dodd, R. B. and R. J. Read (2009). Structures of two human ubiquitin-conjugating enzymes from twinned crystals.
Domazet-Loso, T. and D. Tautz (2010). "Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa." BMC Biol **8**: 66.
Dong, Y., M. Xie, Y. Jiang, et al. (2013). "Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus)." Nat Biotechnol **31**(2): 135-141.
Dorus, S., E. J. Vallender, P. D. Evans, et al. (2004). "Accelerated evolution of nervous system genes in the origin of Homo sapiens." Cell **119**(7): 1027-1040.
Drake, J. W., B. Charlesworth, D. Charlesworth, et al. (1998). "Rates of spontaneous mutation." Genetics **148**(4): 1667-1686.
Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, et al. (2002). "Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data." Genetics **161**(3): 1307-1320.
Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.
Effron, M., L. A. Griner and K. Benirschke (1977). "Nature and rate of neoplasia found in captive wild mammals, birds and reptiles at necropsy." J Natl. Cancer Inst. **59**(1): 185-198.

Eizirik, E., W. J. Murphy and S. J. O'Brien (2001). "Molecular dating and biogeography of the early placental mammal radiation." J Hered **92**(2): 212-219.
English, A. C., S. Richards, Y. Han, et al. (2012). "Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology." PLoS One **7**(11): e47768.
Eswar, N., B. John, N. Mirkovic, et al. (2003). "Tools for comparative protein structure modeling and analysis." Nucleic Acids Res **31**(13): 3375-3380.
Etzioni, R., N. Urban, S. Ramsey, et al. (2003). "The case for early detection." Nature Reviews Cancer **3**(4): 243-252.
Felsenstein, J. (2003). Inferring phylogenies. Sunderland, Mass.,, Sinauer Associates.
Ferlay, J., H. R. Shin, F. Bray, et al. (2010). "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008." Int J Cancer **127**(12): 2893-2917.
Franceschini, A., D. Szklarczyk, S. Frankild, et al. (2013). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." Nucleic Acids Res **41**(Database issue): D808-815.
Frank, S. A. (2007). Dynamics of cancer : incidence, inheritance, and evolution. Princeton, N.J., Princeton University Press.
Fredrickson, E. K. and R. G. Gardner (2012). "Selective destruction of abnormal proteins by ubiquitin-mediated protein quality control degradation." Semin Cell Dev Biol **23**(5): 530-537.
García-Cao, I., M. García-Cao, J. Martín-Caballero, et al. (2002). ""Super p53" mice exhibit enhanced DNA damage response, are tumor resistant and age normally." The EMBO journal **21**(22): 6225-6235.
Garcia-Cao, I., M. Garcia-Cao, J. Martin-Caballero, et al. (2002). ""Super p53" mice exhibit enhanced DNA damage response, are tumor resistant and age normally." The EMBO journal **21**(22): 6225-6235.
Garland, T., Jr., A. F. Bennett and E. L. Rezende (2005). "Phylogenetic approaches in comparative physiology." J Exp Biol **208**(Pt 16): 3015-3035.
Gatenby, R., R. Gillies and J. Brown (2010). "The evolutionary dynamics of cancer prevention." Nature Reviews Cancer **10**(8): 526-527.
Gearing, D. P., N. M. Gough, J. A. King, et al. (1987). "Molecular cloning and expression of cDNA encoding a murine myeloid leukaemia inhibitory factor (LIF)." EMBO J **6**(13): 3995-4002.
Giaccia, A. J. and M. B. Kastan (1998). "The complexity of p53 modulation: emerging patterns from divergent signals." Genes Dev **12**(19): 2973-2983.
Gibbs, R. A., J. Rogers, M. G. Katze, et al. (2007). "Evolutionary and biomedical insights from the rhesus macaque genome." Science **316**(5822): 222-234.
Gilmore, T. D. (2006). "Introduction to NF-kappaB: players, pathways, perspectives." Oncogene **25**(51): 6680-6684.
Gnerre, S., I. Maccallum, D. Przybylski, et al. (2011). "High-quality draft assemblies of mammalian genomes from massively parallel sequence data." Proc Natl Acad Sci U S A **108**(4): 1513-1518.
Gonzalez, K. D., K. A. Noltner, C. H. Buzin, et al. (2009). "Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations." J Clin Oncol **27**(8): 1250-1256.
Gouy, M., S. Guindon and O. Gascuel (2010). "SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building." Mol Biol Evol **27**(2): 221-224.
Grabherr, M. G., B. J. Haas, M. Yassour, et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol **29**(7): 644-652.

Graham, J. (1992). Cancer Selection: The New Theory of Evolution. USA, Aculeus Press.
Greaves, M. (2007). "Darwinian medicine: a case for cancer." Nat Rev Cancer **7**(3): 213-221.
Greenman, C., P. Stephens, R. Smith, et al. (2007). "Patterns of somatic mutation in human cancer genomes." Nature **446**(7132): 153-158.
Gregory, T. R. (2013). Animal Genome Size Database.
Griner, L. A. (1983). Pathology of Zoo Animals. San Diego, CA, Zoological Society of San Diego.
Grunbaum, U., A. Meye, M. Bache, et al. (2001). "Transfection with mdm2-antisense or wtp53 results in radiosensitization and an increased apoptosis of a soft tissue sarcoma cell line." Anticancer Res **21**(3B): 2065-2071.
Guiler, E. R. (1983). Tasmanian Devil. The Australian Museum Complete Book of Australian Mammals. R. Strahan, Anugus & Robertson.
Guindon, S., J. F. Dufayard, V. Lefort, et al. (2010). "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0." Syst Biol **59**(3): 307-321.
Hahn, M. A., K. A. Dickson, S. Jackson, et al. (2012). "The tumor suppressor CDC73 interacts with the ring finger proteins RNF20 and RNF40 and is required for the maintenance of histone 2B monoubiquitination." Hum Mol Genet **21**(3): 559-568.
Hahn, W. and R. Weinberg (2002). "Rules for making human tumor cells." The New England journal of medicine **347**(20): 1593-1603.
Hanahan, D. (2000). "The Hallmarks of Cancer." Cell **100**(1): 57-70.
Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.
Herman, A. B., V. M. Savage and G. B. West (2011). "A quantitative theory of solid tumor growth, metabolic rate and vascularization." PLoS One **6**(9): e22973.
Hershko, A., H. Heller, S. Elias, et al. (1983). "Components of ubiquitin-protein ligase system. Resolution, affinity purification, and role in protein breakdown." J Biol Chem **258**(13): 8206-8214.
Heyer, B. S., A. MacAuley, O. Behrendtsen, et al. (2000). "Hypersensitivity to DNA damage leads to increased apoptosis during early mouse development." Genes Dev **14**(16): 2072-2084.
Higgins, M. E., M. Claremont, J. E. Major, et al. (2007). "CancerGenes: a gene selection resource for cancer genome projects." Nucleic Acids Res **35**(Database issue): D721-726.
Hirotsune, S., N. Yoshida, A. Chen, et al. (2003). "An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene." Nature **423**(6935): 91-96.
Hoeijmakers, J. H. (2009). "DNA damage, aging, and cancer." N Engl J Med **361**(15): 1475-1485.
Holliseter-Smith, J. A., J. H. Poole, E. A. Archie, et al. (2007). "Age, musth and paternity success in wild male African elephants, Loxodonta africana." Animal Behaviour **74**: 287-296.
Hollstein, M., D. Sidransky, B. Vogelstein, et al. (1991). "p53 mutations in human cancers." Science **253**(5015): 49-53.
Horne, H. N., P. S. Lee, S. K. Murphy, et al. (2009). "Inactivation of the MAL gene in breast cancer is a common event that predicts benefit from adjuvant chemotherapy." Mol Cancer Res **7**(2): 199-209.
Horton, K. M., F. M. Corl and E. K. Fishman (2000). "CT evaluation of the colon: inflammatory disease." Radiographics **20**(2): 399-418.

Hoyert, D. and J. Xu (2012). "Deaths: Preliminary Data for 2011." National Vital Statistics Reports **61**(6).
Huang da, W., B. T. Sherman and R. A. Lempicki (2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic Acids Res **37**(1): 1-13.
Huang da, W., B. T. Sherman and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.
Humbert, O., S. Fiumicino, G. Aquilina, et al. (1999). "Mismatch repair and differential sensitivity of mouse and human cells to methylating agents." Carcinogenesis **20**(2): 205-214.
Ise, K., K. Nakamura, K. Nakao, et al. (2000). "Targeted deletion of the H-ras gene decreases tumor formation in mouse skin carcinogenesis." Oncogene **19**(26): 2951-2956.
Issaeva, N., P. Bozko, M. Enge, et al. (2004). "Small molecule RITA binds to p53, blocks p53-HDM-2 interaction and activates p53 function in tumors." Nat Med **10**(12): 1321-1328.
Ivanova, D. G. and T. M. Yankova (2013). "The free radical theory of aging in search of a strategy for increasing life span." Folia Med (Plovdiv) **55**(1): 33-41.
IWC (2013). Status of Whales. iwc.int/status, International Whaling Commission.
Jackson, J. A., C. S. Baker, M. Vant, et al. (2009). "Big and slow: phylogenetic estimates of molecular evolution in baleen whales (suborder mysticeti)." Mol Biol Evol **26**(11): 2427-2440.
Ji, L. and J. A. Roth (2008). "Tumor suppressor FUS1 signaling pathway." J Thorac Oncol **3**(4): 327-330.
Johnsson, P., A. Ackley, L. Vidarsdottir, et al. (2013). "A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells." Nat Struct Mol Biol **20**(4): 440-446.
Jones, S., W. D. Chen, G. Parmigiani, et al. (2008). "Comparative lesion sequencing provides insights into tumor evolution." Proc Natl Acad Sci U S A **105**(11): 4283-4288.
Juang, Y. C., M. C. Landry, M. Sanches, et al. (2012). "OTUB1 co-opts Lys48-linked ubiquitin recognition to suppress E2 enzyme function." Mol Cell **45**(3): 384-397.
Jung, K. H., J. H. Noh, J. K. Kim, et al. (2012). "HDAC2 overexpression confers oncogenic potential to human lung cancer cells by deregulating expression of apoptosis and cell cycle proteins." J Cell Biochem **113**(6): 2167-2177.
Kang, H. and D. Shibata (2013). "Direct Measurements of Human Colon Crypt Stem Cell Niche Genetic Fidelity: The Role of Chance in Non-Darwinian Mutation Selection." Front Oncol **3**: 264.
Kapitonov, V. V., G. P. Holmquist and J. Jurka (1998). "L1 repeat is a basic unit of heterochromatin satellites in cetaceans." Mol Biol Evol **15**(5): 611-612.
Kar, G., O. Keskin, R. Nussinov, et al. (2012). "Human proteome-scale structural modeling of E2-E3 interactions exploiting interface motifs." J Proteome Res **11**(2): 1196-1207.
Karin, M. (2006). "Nuclear factor-kappaB in cancer development and progression." Nature **441**(7092): 431-436.
Kim, D., G. Pertea, C. Trapnell, et al. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome Biol **14**(4): R36.
Kinzler, K. W. and B. Vogelstein (1997). "Cancer-susceptibility genes. Gatekeepers and caretakers." Nature **386**(6627): 761, 763.
Kirkwood, T. B. (2005). "Understanding the odd science of aging." Cell **120**(4): 437-447.
Kitagawa, K., Y. Kotake and M. Kitagawa (2009). "Ubiquitin-mediated control of oncogene and tumor suppressor gene products." Cancer Sci **100**(8): 1374-1381.

Kitazoe, Y., H. Kishino, P. J. Waddell, et al. (2007). "Robust time estimation reconciles views of the antiquity of placental mammals." PLoS One **2**(4): e384.
Klein, G. (2009). "Toward a genetics of cancer resistance." Proceedings of the National Academy of Sciences of the United States of America **106**(3): 859-863.
Knaub, A. D. (2001). Salt and Friends.
Knaub, A. D. (2012) "A Whale in Danger."
Knoll, A. H., E. J. Javaux, D. Hewitt, et al. (2006). "Eukaryotic organisms in Proterozoic oceans." Philos Trans R Soc Lond B Biol Sci **361**(1470): 1023-1038.
Knudson, A. G., Jr. (1971). "Mutation and cancer: statistical study of retinoblastoma." Proc Natl Acad Sci U S A **68**(4): 820-823.
Ko, L. J. and C. Prives (1996). "p53: puzzle and paradigm." Genes Dev **10**(9): 1054-1072.
Koebel, C., W. Vermi, J. Swann, et al. (2007). "Adaptive immunity maintains occult cancer in an equilibrium state." Nature **450**(7171): 903-907.
Koehl, D. (1995-2012). Elephant Encyclopedia.
Kohany, O., A. J. Gentles, L. Hankus, et al. (2006). "Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor." BMC Bioinformatics **7**: 474.
Komarova, E. A., K. Christov, A. I. Faerman, et al. (2000). "Different impact of p53 and p21 on the radiation response of mouse tissues." Oncogene **19**(33): 3791-3798.
Kondrashov, A. S. (2003). "Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases." Hum Mutat **21**(1): 12-27.
Ku, H. H., U. T. Brunk and R. S. Sohal (1993). "Relationship between mitochondrial superoxide and hydrogen peroxide production and longevity of mammalian species." Free Radic Biol Med **15**(6): 621-627.
Kundu, K., S. F. Knight, N. Willett, et al. (2009). "Hydrocyanines: a class of fluorescent sensors that can image reactive oxygen species in cell culture, tissue, and in vivo." Angew Chem Int Ed Engl **48**(2): 299-303.
Lakin, N. D. and S. P. Jackson (1999). "Regulation of p53 in response to DNA damage." Oncogene **18**(53): 7644-7655.
Lambersten, R. H., C. S. Baker, D. A. Duffield, et al. (1988). "Cytogenetic determination of sex among individually identified humpback whales (Megaptera novaeangliae)." Canadian Journal of Zoology **66**: 1243-1248.
Lambertsen, R. H. (1987). "A Biopsy System for Large Whales and Its Use for Cytogenetics." Jounal of Mammalogy **68**(2): 443-445.
Lander, E. S., L. M. Linton, B. Birren, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.
Lane, D. P. (1992). "Cancer. p53, guardian of the genome." Nature **358**(6381): 15-16.
Leroi, A., V. Koufopanou and A. Burt (2003). "Cancer selection." Nature reviews. Cancer **3**(3): 226-231.
Levy, S., G. Sutton, P. C. Ng, et al. (2007). "The diploid genome sequence of an individual human." PLoS Biol **5**(10): e254.
Li, R., W. Fan, G. Tian, et al. (2010). "The sequence and de novo assembly of the giant panda genome." Nature **463**(7279): 311-317.
Li, X. and Y. H. Chang (1995). "Amino-terminal protein processing in Saccharomyces cerevisiae is an essential function that requires two distinct methionine aminopeptidases." Proc Natl Acad Sci U S A **92**(26): 12357-12361.
Li, Y. and J. P. de Magalhaes (2013). "Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity." Age **35**(2): 301-314.

Li, Y., J. He, F. Wang, et al. (2010). "Differentiation of embryonic stem cells in adult bone marrow." J Genet Genomics **37**(7): 431-439.
Liang, X. H., S. Jackson, M. Seaman, et al. (1999). "Induction of autophagy and inhibition of tumorigenesis by beclin 1." Nature **402**(6762): 672-676.
Lichtenstein, A. (2005). "On evolutionary origin of cancer." Cancer Cell International **5**(1): 5.
Lichtenstein, A. V. (2005). "Cancer as a programmed death of an organism." Biochemistry (Mosc) **70**(9): 1055-1064.
Lind, G. E., T. Ahlquist, M. Kolberg, et al. (2008). "Hypermethylated MAL gene - a silent marker of early colon tumorigenesis." J Transl Med **6**: 13.
Liu, Y., J. A. Cotton, B. Shen, et al. (2010). "Convergent sequence evolution between echolocating bats and dolphins." Current biology : CB **20**(2): R53-54.
Loeb, L. A. (1991). "Mutator phenotype may be required for multistage carcinogenesis." Cancer Res **51**(12): 3075-3079.
Longo, V. D. and L. Fontana (2010). "Calorie restriction and cancer prevention: metabolic and molecular mechanisms." Trends Pharmacol Sci **31**(2): 89-98.
Loytynoja, A. and N. Goldman (2008). "Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis." Science **320**(5883): 1632-1635.
Lynch, M. (2010). "Rate, molecular spectrum, and consequences of human mutation." Proc Natl Acad Sci U S A **107**(3): 961-968.
Mallery, D. L., C. J. Vandenberg and K. Hiom (2002). "Activation of the E3 ligase function of the BRCA1/BARD1 complex by polyubiquitin chains." EMBO J **21**(24): 6755-6762.
Mardis, E. R., L. Ding, D. J. Dooling, et al. (2009). "Recurring mutations found by sequencing an acute myeloid leukemia genome." N Engl J Med **361**(11): 1058-1066.
Marsh, H. (1986). "Evidence for reproductive senescence in female cetaceans." Rep. Int. Whale Commn.(8): 57-73.
Martens, E. A., R. Kostadinov, C. C. Maley, et al. (2011). "Spatial structure increases the waiting time for cancer." New J Phys **13**.
Martineau, D., K. Lemberger, A. Dallaire, et al. (2002). "Cancer in wildlife, a case study: beluga from the St. Lawrence estuary, Québec, Canada." Environmental health perspectives **110**(3): 285-292.
Matheu, A., C. Pantoja, A. Efeyan, et al. (2004). "Increased gene dosage of Ink4a/Arf results in cancer resistance and normal aging." Genes & development **18**(22): 2736-2746.
McAloose, D. and A. L. Newton (2009). "Wildlife cancer: a conservation perspective." Nat Rev Cancer **9**(7): 517-526.
McComb, K., G. Shannon, S. M. Durant, et al. (2011). "Leadership in elephants: the adaptive value of age." Proc Biol Sci **278**(1722): 3270-3276.
Merlo, L., J. Pepper, B. Reid, et al. (2006). "Cancer as an evolutionary and ecological process." Nature Reviews Cancer **6**(12): 924-935.
Michor, F. (2007). "Chronic myeloid leukemia blast crisis arises from progenitors." Stem Cells **25**(5): 1114-1118.
Miller, J. R., S. Koren and G. Sutton (2010). "Assembly algorithms for next-generation sequencing data." Genomics **95**(6): 315-327.
Mimori, K., T. Shiraishi, K. Mashino, et al. (2003). "MAL gene expression in esophageal cancer suppresses motility, invasion and tumorigenicity and enhances apoptosis through the Fas pathway." Oncogene **22**(22): 3463-3471.
Minor, L. K. (2008). "Label-free cell-based functional assays." Comb Chem High Throughput Screen **11**(7): 573-580.

Miyashita, T., S. Krajewski, M. Krajewska, et al. (1994). "Tumor suppressor p53 is a regulator of bcl-2 and bax gene expression in vitro and in vivo." Oncogene **9**(6): 1799-1805.
Miyata, T. and T. Yasunaga (1980). "Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application." J Mol Evol **16**(1): 23-36.
Monaghan, P. (2010). "Telomeres and life histories: the long and the short of it." Ann N Y Acad Sci **1206**(1): 130-142.
Morris, J. and J. Dobson (2001). Small Animal Oncology, Wiley-Blackwell.
Morris, J. R., L. Pangon, C. Boutell, et al. (2006). "Genetic analysis of BRCA1 ubiquitin ligase activity and its relationship to breast cancer susceptibility." Hum Mol Genet **15**(4): 599-606.
Morris, J. R. and E. Solomon (2004). "BRCA1 : BARD1 induces the formation of conjugated ubiquitin structures, dependent on K6 of ubiquitin, in cells during DNA replication and repair." Hum Mol Genet **13**(8): 807-817.
Murphy, W. J., T. H. Pringle, T. A. Crider, et al. (2007). "Using genomic data to unravel the root of the placental mammal phylogeny." Genome Res **17**(4): 413-421.
Myers, E. W., G. G. Sutton, A. L. Delcher, et al. (2000). "A whole-genome assembly of Drosophila." Science **287**(5461): 2196-2204.
Nagy, J. D., E. M. Victor and J. H. Cropper (2007). "Why don't all whales have cancer? A novel hypothesis resolving Peto's paradox." Integrative and Comparative Biology **47**(2): 317-328.
Newman, D. J. and G. M. Cragg (2007). "Natural products as sources of new drugs over the last 25 years." J Nat Prod **70**(3): 461-477.
Newman, S. J. and S. A. Smith (2006). "Marine mammal neoplasia: a review." Vet Pathol **43**(6): 865-880.
NOAA (2006). Salt. http://stellwagen.noaa.gov/visit/whalewatching/top50/salt.html, Gerry E. Studds Stellwagen Bank National Marine Sanctuary.
Nowell, P. C. (1976). "The clonal evolution of tumor cell populations." Science **194**(4260): 23-28.
Nunney, L. (1999). "Lineage Selection and the Evolution of Multistage Carcinogenesis." Proceedings: Biological Sciences **266**(1418).
Nunney, L. (2013). "The real war on cancer: the evolutionary dynamics of cancer suppression." Evol Appl **6**(1): 11-19.
Ohta, T. (1973). "Slightly deleterious mutant substitutions in evolution." Nature **246**(5428): 96-98.
Olive, P. L. and J. P. Banath (2006). "The comet assay: a method to measure DNA damage in individual cells." Nat Protoc **1**(1): 23-29.
Palsboll, P., F. Larsen and E. Hensen (1991). Sampling of skin biopsies from free-ranging large cetaceans in West Greenland: Development of new biopsy tips and bolt designs, International Whaling Commission. **13:** 71-79.
Parra, G., K. Bradnam, Z. Ning, et al. (2009). "Assessing the gene space in draft genomes." Nucleic Acids Res **37**(1): 289-297.
Patterson, N., D. J. Richter, S. Gnerre, et al. (2006). "Genetic evidence for complex speciation of humans and chimpanzees." Nature **441**(7097): 1103-1108.
Pawelec, G., E. Derhovanessian and A. Larbi (2010). "Immunosenescence and cancer." Critical Reviews in Oncology/Hematology **75**(2): 165-172.

Payne, S. and C. Kemp (2005). "Tumor suppressor genetics." Carcinogenesis **26**(12): 2031-2045.
Pearson, B. J. and A. Sánchez Alvarado (2008). "Regeneration, stem cells, and the evolution of tumor suppression." Cold Spring Harbor symposia on quantitative biology **73**: 565-572.
Pei, B., C. Sisu, A. Frankish, et al. (2012). "The GENCODE pseudogene resource." Genome Biol **13**(9): R51.
Pepper, J., K. Sprouffske and C. Maley (2007). "Animal cell differentiation patterns suppress somatic evolution." PLoS computational biology **3**(12): e250.
Perry, M. E. (2004). "Mdm2 in the response to radiation." Mol Cancer Res **2**(1): 9-19.
Peto, R. (1977). Epidemiology, multistage models, and short-term mutagenicity tests. The Origins of Human Cancer, Cold Spring Harbor Conferences on Cell Proliferation, Cold Spring Harbor Laboratory.
Peto, R., F. J. C. Roe, P. N. Lee, et al. (1975). "Cancer and Aging in Mice and Men." British Journal of Cancer **32**(4): 411-426.
Phillippy, A. M., M. C. Schatz and M. Pop (2008). "Genome assembly forensics: finding the elusive mis-assembly." Genome Biol **9**(3): R55.
Pickhardt, P. J., R. B. Halberg, A. J. Taylor, et al. (2005). "Microcomputed tomography colonography for polyp detection in an in vivo mouse tumor model." Proc Natl Acad Sci U S A **102**(9): 3419-3422.
Pink, R. C., K. Wicks, D. P. Caley, et al. (2011). "Pseudogenes: pseudo-functional or key regulators in health and disease?" RNA **17**(5): 792-798.
Poliseno, L., L. Salmena, J. Zhang, et al. (2010). "A coding-independent function of gene and pseudogene mRNAs regulates tumour biology." Nature **465**(7301): 1033-1038.
Pollier, J., S. Rombauts and A. Goossens (2013). "Analysis of RNA-Seq data with TopHat and Cufflinks for genome-wide expression analysis of jasmonate-treated plants and plant cultures." Methods Mol Biol **1011**: 305-315.
Pop, M. and S. L. Salzberg (2008). "Bioinformatics challenges of new sequencing technology." Trends Genet **24**(3): 142-149.
Prabhakar, S., J. P. Noonan, S. Paabo, et al. (2006). "Accelerated evolution of conserved noncoding sequences in humans." Science **314**(5800): 786.
Puertollano, R. and M. A. Alonso (1999). "MAL, an integral element of the apical sorting machinery, is an itinerant protein that cycles between the trans-Golgi network and the plasma membrane." Mol Biol Cell **10**(10): 3435-3447.
Qu, X., J. Yu, G. Bhagat, et al. (2003). "Promotion of tumorigenesis by heterozygous disruption of the beclin 1 autophagy gene." J Clin Invest **112**(12): 1809-1820.
Rangarajan, A., S. J. Hong, A. Gifford, et al. (2004). "Species- and cell type-specific requirements for cellular transformation." Cancer Cell **6**(2): 171-183.
Rangarajan, A. and R. A. Weinberg (2003). "Opinion: Comparative biology of mouse versus human cells: modelling human cancer in mice." Nature Reviews Cancer **3**(12): 952-959.
Rattan, R., K. Narita, J. Chien, et al. (2010). "TCEAL7, a putative tumor suppressor gene, negatively regulates NF-kappaB pathway." Oncogene **29**(9): 1362-1373.
Ribble, D., N. B. Goldstein, D. A. Norris, et al. (2005). "A simple technique for quantifying apoptosis in 96-well plates." BMC Biotechnol **5**: 12.
Rice, P., I. Longden and A. Bleasby (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-277.
Riggins, G. J. and R. L. Strausberg (2001). "Genome and genetic resources from the Cancer Genome Anatomy Project." Hum Mol Genet **10**(7): 663-667.

Rippin, T. M., V. J. Bykov, S. M. Freund, et al. (2002). "Characterization of the p53-rescue drug CP-31398 in vitro and in living cells." Oncogene **21**(14): 2119-2129.
Roche, B., M. E. Hochberg, A. F. Caulin, et al. (2012). "Natural resistance to cancers: a Darwinian hypothesis to explain Peto's paradox." BMC cancer **12**: 387.
Roche, B., K. Sprouffske, H. Hbid, et al. (2013). "Peto's paradox revisited: theoretical evolutionary dynamics of cancer in wild populations." Evol Appl **6**(1): 109-116.
Ropero, S., E. Ballestar, M. Alaminos, et al. (2008). "Transforming pathways unleashed by a HDAC2 mutation in human cancer." Oncogene **27**(28): 4008-4012.
Rost, B. and J. Liu (2003). "The PredictProtein server." Nucleic Acids Res **31**(13): 3300-3304.
Rost, B., G. Yachdav and J. Liu (2004). "The PredictProtein server." Nucleic Acids Res **32**(Web Server issue): W321-326.
Rothschild, B. M., D. H. Tanke, M. Helbling, 2nd, et al. (2003). "Epidemiologic study of tumors in dinosaurs." Naturwissenschaften **90**(11): 495-500.
Santra, M. K., N. Wajapeyee and M. R. Green (2009). "F-box protein FBXO31 mediates cyclin D1 degradation to induce G1 arrest after DNA damage." Nature **459**(7247): 722-725.
Savage, V. M., A. P. Allen, J. H. Brown, et al. (2007). "Scaling of number, size, and metabolic rate of cells with body size in mammals." Proc Natl Acad Sci U S A **104**(11): 4718-4723.
Schmidly, D. (1994). The Mammals of Texas. Texas, University of Texas Press.
Schrodinger, LLC (2013). The PyMOL Molecular Graphics System, Version 1.6.0.0.
Sedelnikova, O. A., C. E. Redon, J. S. Dickey, et al. (2010). "Role of oxidatively induced DNA lesions in human pathogenesis." Mutat Res **704**(1-3): 152-159.
SEER (2001). Surveillance, Epidemiology and End Results (SEER) Program: SEER*Stat Database: Incidence - SEER 11 Regs Public-Use. http://www.seer.cancer.gov/, Sureveillance Research Program, Cancer Statistics Branch.
Seim, I., X. Fang, Z. Xiong, et al. (2013). "Genome analysis reveals insights into physiology and longevity of the Brandt's bat Myotis brandtii." Nat Commun **4**: 2212.
Seluanov, A., Z. Chen, C. Hine, et al. (2007). "Telomerase activity coevolves with body mass not lifespan." Aging cell **6**(1): 45-52.
Seluanov, A., C. Hine, J. Azpurua, et al. (2009). "Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat." Proceedings of the National Academy of Sciences of the United States of America **106**(46): 19352-19357.
Seluanov, A., C. Hine, M. Bozzella, et al. (2008). "Distinct tumor suppressor mechanisms evolve in rodent species that differ in size and lifespan." Aging cell **7**(6): 813-823.
Shangary, S., D. Qin, D. McEachern, et al. (2008). "Temporal activation of p53 by a specific MDM2 inhibitor is selectively toxic to tumors and leads to complete tumor growth inhibition." Proc Natl Acad Sci U S A **105**(10): 3933-3938.
Shankaran, V., H. Ikeda, A. T. Bruce, et al. (2001). "IFNgamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity." Nature **410**(6832): 1107-1111.
Shay, J. W. and W. E. Wright (2010). "Telomeres and telomerase in normal and cancer stem cells." FEBS Lett **584**(17): 3819-3825.
She, R., J. S. Chu, B. Uyar, et al. (2011). "genBlastG: using BLAST searches to build homologous gene models." Bioinformatics **27**(15): 2141-2143.
She, R., J. S. Chu, K. Wang, et al. (2009). "GenBlastA: enabling BLAST to identify homologous gene sequences." Genome Res **19**(1): 143-149.

Siegal, F. P., N. Kadowaki, M. Shodell, et al. (1999). "The nature of the principal type 1 interferon-producing cells in human blood." Science **284**(5421): 1835-1837.
Simon, A. (2012). FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
Simpson, J. T. and R. Durbin (2012). "Efficient de novo assembly of large genomes using compressed data structures." Genome Res **22**(3): 549-556.
Singh, N. P., C. H. Muller and R. E. Berger (2003). "Effects of age on DNA double-strand breaks and apoptosis in human sperm." Fertil Steril **80**(6): 1420-1430.
Sjoblom, T., S. Jones, L. D. Wood, et al. (2006). "The consensus coding sequences of human breast and colorectal cancers." Science **314**(5797): 268-274.
Small, G. L. (1971). The blue whale. New York,, Columbia University Press.
Smith, F. A. (2003). Body mass of late Quaternary mammals. Ecology. **84:** 3403.
Snippert, H. J., L. G. van der Flier, T. Sato, et al. (2010). "Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells." Cell **143**(1): 134-144.
Soissi, T. (1996). "The TP53 Web Site." from p53.free.fr.
Solomon, H., S. Madar and V. Rotter (2011). "Mutant p53 gain of function is interwoven into the hallmarks of cancer." The Journal of pathology **225**(4): 475-478.
Spaulding, C. C., R. L. Walford and R. B. Effros (1997). "The accumulation of non-replicative, non-functional, senescent T cells with age is avoided in calorically restricted mice by an enhancement of T cell apoptosis." Mechanisms of ageing and development **93**(1-3): 25-33.
Speakman, J. R. (2005). "Body size, energy metabolism and lifespan." J Exp Biol **208**(Pt 9): 1717-1730.
Sreedhar, A. (2010). "P53 pseudogene: potential role in heat shock induced apoptosis in a rat histiocytoma." Health(2): 1065-1071.
Stadtman, E. R. (2004). "Role of oxidant species in aging." Curr Med Chem **11**(9): 1105-1112.
Stajich, J. E., D. Block, K. Boulez, et al. (2002). "The Bioperl toolkit: Perl modules for the life sciences." Genome Res **12**(10): 1611-1618.
Strauss, B. S. (1992). "The origin of point mutations in human tumor cells." Cancer Res **52**(2): 249-253.
Takaoka, A., S. Hayakawa, H. Yanai, et al. (2003). "Integration of interferon-alpha/beta signalling to p53 responses in tumour suppression and antiviral defence." Nature **424**(6948): 516-523.
Talia, P., E. J. Greizerstein, H. E. Hopp, et al. (2011). "Detection of single copy sequences using BAC-FISH and C-PRINS techniques in sunflower chromosomes." Biocell **35**(1): 19-28.
Tariq, M. A., H. J. Kim, O. Jejelowo, et al. (2011). "Whole-transcriptome RNAseq analysis from minute amount of total RNA." Nucleic Acids Res **39**(18): e120.
Teilhard de Chardin, P. (1959). The phenomenon of man. New York,, Harper.
Testa, J. R., D. Malkin and J. D. Schiffman (2013). "Connecting molecular pathways to hereditary cancer risk syndromes." Am Soc Clin Oncol Educ Book **2013**: 81-90.
Tokino, T. and Y. Nakamura (2000). "The role of p53-target genes in human cancer." Crit Rev Oncol Hematol **33**(1): 1-6.
Totter, J. R. (1980). "Spontaneous cancer and its possible relationship to oxygen metabolism." Proc Natl Acad Sci U S A **77**(4): 1763-1767.
Treangen, T. J. and S. L. Salzberg (2012). "Repetitive DNA and next-generation sequencing: computational challenges and solutions." Nat Rev Genet **13**(1): 36-46.

Turturro, A., W. W. Witt, S. Lewis, et al. (1999). "Growth curves and survival characteristics of the animals used in the Biomarkers of Aging Program." J Gerontol A Biol Sci Med Sci **54**(11): B492-501.
Tyner, S. D., S. Venkatachalam, J. Choi, et al. (2002). "p53 mutant mice that display early ageing-associated phenotypes." Nature **415**(6867): 45-53.
van Meerbeeck, P. (1979). "[Image of body and family body (author's transl)]." Acta Psychiatr Belg **79**(6): 614-622.
van Wijk, S. J., S. J. de Vries, P. Kemmeren, et al. (2009). "A comprehensive framework of E2-RING E3 interactions of the human ubiquitin-proteasome system." Mol Syst Biol **5**: 295.
Vassilev, L. T., B. T. Vu, B. Graves, et al. (2004). "In vivo activation of the p53 pathway by small-molecule antagonists of MDM2." Science **303**(5659): 844-848.
Vavouri, T., J. I. Semple and B. Lehner (2008). "Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution." Trends Genet **24**(10): 485-488.
Ventura, A., D. G. Kirsch, M. E. McLaughlin, et al. (2007). "Restoration of p53 function leads to tumour regression in vivo." Nature **445**(7128): 661-665.
Ward, E. J., K. Parsons, E. E. Holmes, et al. (2009). "The role of menopause and reproductive senescence in a long-lived social mammal." Front Zool **6**: 4.
Welchman, R. L., C. Gordon and R. J. Mayer (2005). "Ubiquitin and ubiquitin-like proteins as multifunctional signals." Nat Rev Mol Cell Biol **6**(8): 599-609.
Wilson, A. J., E. Holson, F. Wagner, et al. (2011). "The DNA damage mark pH2AX differentiates the cytotoxic effects of small molecule HDAC inhibitors in ovarian cancer cells." Cancer Biol Ther **12**(6): 484-493.
Wiseman, H. and B. Halliwell (1996). "Damage to DNA by reactive oxygen and nitrogen species: role in inflammatory disease and progression to cancer." Biochem J **313 (Pt 1)**: 17-29.
Woosley, J. T. (1991). "Measuring cell proliferation." Arch Pathol Lab Med **115**(6): 555-557.
Xie, Y., H. Yang, C. Cunanan, et al. (2004). "Deficiencies in Mouse Myh and Ogg1 Result in Tumor Predisposition and G to T Mutations in Codon 12 of the K-Ras Oncogene in Lung Tumors." Cancer Research **64**(9): 3096-3102.
Xu, Z., E. Kohli, K. I. Devlin, et al. (2008). "Interactions between the quality control ubiquitin ligase CHIP and ubiquitin conjugating enzymes." BMC Struct Biol **8**: 26.
Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-1591.
Yatabe, Y., S. Tavare and D. Shibata (2001). "Investigating stem cells in human colon by using methylation patterns." Proc Natl Acad Sci U S A **98**(19): 10839-10844.
Ye, Y. and M. Rape (2009). "Building ubiquitin chains: E2 enzymes at work." Nat Rev Mol Cell Biol **10**(11): 755-764.
Yim, H. S., Y. S. Cho, X. Guang, et al. (2013). "Minke whale genome and aquatic adaptation in cetaceans." Nat Genet.
Zelnick, C. R., D. J. Burks and C. H. Duncan (1987). "A composite transposon 3' to the cow fetal globin gene binds a sequence specific factor." Nucleic Acids Res **15**(24): 10437-10453.
Zhang, J. and S. Kumar (1997). "Detection of convergent and parallel evolution at the amino acid sequence level." Mol Biol Evol **14**(5): 527-536.
Zimin, A. V., G. Marcais, D. Puiu, et al. (2013). "The MaSuRCA genome assembler." Bioinformatics.

zur Hausen, H. (1999). "Viral Oncogenesis." JAIDS Journal of Acquired Immune Deficiency Syndromes **21**(1): A7.